Jingyi Zhou. An Analysis of the Allergy Comments on Twitter Using Data Mining Approach. A Master's Paper for the M.S. in I.S degree. April, 2020. 60 pages. Advisor: Jaime Arguello

Allergies are one of the most common chronic illnesses in the world. The prevalence of social media allows people to express their opinions and exchange information including symptoms of personal health. Mining those publicly accessible health-related data on social media, such as Twitter, offers a unique approach to get valuable healthcare insights.

In this paper, a multi-component data mining framework was developed to collect Twitter data, detect time series patterns, discover topics of interest about allergies, and analyze the contents of tweets. From the extracted 2.2 million tweets in 2019, my experimental results show that allergy-related tweet volume is strongly correlated to the pollen data ($r = .699$, $p < .01$). Also, 152 unique topics are identified with a -28.36 perplexity score and a .67 coherence score. Furthermore, many linguistic dimensions such as the sentiment are analyzed to learn about the tweet contents. I consider this to be one of the many studies examining a large-scale social media stream to deeply analyze allergy activities. And with the growing social media, publicly available data such as Twitter posts can be used to support healthcare practitioners and social scientists in better understanding common public opinions, not just allergies.

Headings:

Text Mining

Health Informatics

Social Media

AN ANALYSIS OF THE ALLERGY COMMENTS ON TWITTER USING DATA MINING APPROACH

by
Jingyi Zhou

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2020

Approved by

_____

Advisor's Firstname Lastname

Table of Contents

Introduction

Allergies, also known as allergic diseases, are conditions caused by hypersensitivity of the immune system to typically harmless substances in the environment (McConnell, 2007). Allergies are common and the complexity and severity of allergic diseases are increasing worldwide. In the developed world, about 20% of people are affected by allergic rhinitis, about 6% of people have at least one food allergy, and about 20% are afflicted with atopic dermatitis at some point in time (Wikipedia, 2019). Worldwide, the rise in the prevalence of allergic diseases has continued in the industrialized world for more than 50 years (WHO, 2012). More than 50 million Americans have experienced various types of allergies each year. And Allergies are the 6th leading cause of chronic illness in the U.S. with an annual cost in excess of $18 billion (Acaai, 2019). People visit the emergency room about 200,000 times each year because of food allergies. In 2017, 8.1 percent of adults (19.9 million adults) and 7.7 percent of children (5.6 children) were diagnosed with hay fever (SHST, 2017).

Moreover, continuous use of allergy medication can worsen patients' health conditions and lead to side effects like dizziness, nausea and vomiting, and other serious medical complications. Furthermore, an increasing number of allergy patients gives rise to allergy-related health care costs and leads to reduced work productivity. Asthma-related medical expenses are estimated to cost the U.S. health-care system $82 billion a year, according to a study by U.S. Centers for Disease Control and Prevention. And researchers reporting in the Journal of the American Medical Association states that the costs of food

allergies, from medical care to food to pharmaceuticals are $4,184 per child per year, costing our economy $25 billion, including lost productivity (Robynobrien, 2019). To this end, 4 million workdays are lost due to hay fever alone each year. Therefore, accurate allergy surveillance and forecast is important to minimize the healthcare cost and maximize work productivity loss due to allergy symptoms.

The growth of social media has provided a research opportunity to track public behaviors, information, and opinions about common health issues including allergy. It is estimated that the number of social media users will increase from 2.14 billion in 2015 to 2.95 billion in 2020 (Statista, 2019). Twitter, one of the largest social networking website, allows users to post short text messages called tweets that can be up to 280 characters in length. As of 2018, Twitter had more than 321 million monthly active users. Twitter has been used as a valuable real-time information resource for various applications. On Twitter, people not only make general chatters but also share photos, news, opinions, emotions, and even health conditions including symptoms and medications they are taking for their diseases. In recent years, many researchers have investigated using twitter for disease surveillance, especially for influenza epidemic detection and prediction. Twitter thus provides a unique opportunity to understand users' opinions with respect to the most common health issues (Mejova, Weber, & Macy, 2015). Publicly available Twitter posts have facilitated data collection and leveraged the research at the intersection of public health and data science; thus, informing the research community of major opinions and topics of interest among the general population (Nasukawa&Yi, 2003) that cannot otherwise be collected through traditional means of research (e.g., surveys, interviews, focus groups) (Eichstaedt et al., 2015). Furthermore, analyzing Twitter data can help health

organizations such as state health departments and large healthcare systems to provide health advice and track health opinions of their populations and provide effective health advice when needed (Mejova et al., 2015).

Among computational methods to analyze tweets, computational linguistics is a well-known developed approach to gain insight into a population, track health issues, and discover new knowledge (Paul & Dredze, 2011, 2012). Twitter data has been used for a wide range of health and non-health related applications, such as the stock market (Bollen, Mao, & Zeng, 2011) and election analysis (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Some examples of Twitter data analysis for health-related topics include flu, mental health, Ebola, Zika, medication use, diabetes, weight loss, and obesity.

In my master's paper, I aim to improve public health allergy surveillance on social media and answer three research questions (RQ):

RQ1: What are the trends of the comments related to allergies on Twitter?

RQ2: What are the main topics related to allergies on Twitter?

RQ3: How can the content of the Tweets be analyzed?

I analyze a large scale Twitter data collected over 12 months to monitor the allergy situation and extract some insights. More specifically,

(1) Expository data analysis is employed to find the latent pattern in allergy-related tweets and a time series analysis is used to determine the causality between tweets amount and pollen levels.

(2) To discover topics from the collected tweets, I use a topic modeling approach, Latent Dirichlet Allocation (LDA), that fuzzy clusters the semantically related words into a topic that has an overall theme.

(3) An objective interpretation approach with a lexicon-based approach, Linguistic Inquiry and Word Count (LIWC), is employed to analyze the content of topics.

Related Works

This chapter provides an overview of previous researches on analyzing public health issues (allergy) via social media (Twitter). While this topic has not been studied in-depth, much academic research has been conducted on the issues around it. In order to fully understand this topic, there are three aspects that I think are very important to understand how social media can be used to help analyze public health issues from multiple perspectives. First, I am going to find out the specific topic of health care analysis via social media. The second aspect will explore multiple ways we can deal with social media data, especially in how to handle natural language and unstructured data. Finally, we will inspect the application and implication of the researches.

As we have known, publicly available Twitter posts have facilitated data collection and leveraged the research at the intersection of public health and data science; thus, informing the research community of major opinions and topics of interest among the general population (Nasukawa&Yi, 2003; Wiebe et al., 2003; Zabin & Jefferies, 2008) that cannot otherwise be collected through traditional means of research (e.g., surveys, interviews, focus groups) (Eichstaedt et al., 2015; Wartell, 2015). Furthermore, analyzing Twitter data can help health organizations such as state health departments and large healthcare systems to provide health advice and track health opinions of their populations and provide effective health advice when needed (Mejova et al., 2015).

Among computational methods to analyze tweets, computational linguistics is a well-known developed approach to gain insight into a population, track health issues, and

discover new knowledge (Moreland-Russell, Tabak, Ruhr, & Maier, 2014; Paul & Dredze, 2011, 2012; Zhao et al., 2011). Twitter data has been used for a wide range of health and non-health related applications, such as the stock market (Bollen, Mao, & Zeng, 2011) and election analysis (Tumasjan, Sprenger, Sandner, & Welpe, 2010).

Some examples of Twitter data analysis for health-related topics include flu, mental health, Ebola, Zika, medication use, diabetes, and weight loss and obesity.

To detect influenza epidemics, the traditional methods mostly rely on expensive surveys of hospitals across the country, typically with lag times of one to two weeks for influenza reporting, and even longer for less common diseases (Culotta, 2010). And there have been several recently proposed solutions to estimate a population's health from Internet activity, most notably Google's Flu Trends service, which correlates search term frequency with influenza statistics reported by the Centers for Disease Control and Prevention (CDC). And there are now more possibilities to detect flu due to the prevalence of social media like Twitter. Researchers analyzed messages posted on the micro-blogging site Twitter.com to determine if a similar correlation can be uncovered. They proposed several methods to identify influenza-related messages and compare a number of regression models to correlate these messages with CDC statistics. Using over 500,000 messages spanning 10 weeks, Culotta's team found that their best model achieves a correlation of 0.78 with CDC statistics by leveraging a document classifier to identify relevant messages, which is a very significant breakthrough.

Besides detecting flu, Twitter can also address the challenges of virus outbreak surveillance, such as Zika and Ebola. Zika-related Twitter incidence peaked after the World Health Organization declared an emergency. Five themes were identified from Zika-related

Twitter content (Fu et al., 2016). Fu's team computationally analyzed the contents of 62,547 English Tweets obtained by search API. Topic modeling was used to group bags of words in Tweets into different topics. Although 20 topics were identified using statistical methods, they relied on human judgment to connote them into 5 themes for interpretations. However, using both a statistical algorithm and human curators makes their study relevant to public health. Computational methods assist-not replace-health communicators during emergency responses. Their study highlighted the needs of multilingual Twitter health communication on the Zika virus.

Michelle Odlum and Sunmoo Yoon demonstrated the use of Twitter as a real-time method of Ebola outbreak surveillance to monitor information spread, capture early epidemic detection, and examine the content of public knowledge and attitudes. They collected tweets mentioning Ebola in English during the early stage of the current Ebola outbreak from July 24-August 1, 2014. Their analysis for this observational study includes time series analysis with geologic visualization to observe information dissemination and content analysis using natural language processing to examine public knowledge and attitudes. In a nutshell, a total of 42,236 tweets (16,499 unique and 25,737 retweets) mentioning Ebola were posted and disseminated to 9,362,267,048 people, 63 times higher than the initial number (Odlum & Yoon), 2015. Tweets started to rise in Nigeria 3-7 days prior to the official announcement of the first probable Ebola case. The topics discussed in tweets include risk factors, prevention education, disease trends, and compassion. Because of the analysis of a unique Twitter dataset captured in the early stage of the current Ebola outbreak, their results provide insight into the intersection of social media and public health outbreak surveillance. Their findings demonstrate the usefulness of Twitter mining to

inform public health education. Allison J. Lazard et al. did a research on a similar topic. They conducted a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. And they concluded that social media text mining provides a valuable tool that can be used quickly and efficiently to improve public health communication efforts by collecting and identifying prevalent themes of public concern (Lazard, Scheinfeld, Bernhardt, Wilcox, & Suran, 2015).

Tweets can also augment a public health program that studies emerging patterns of illicit drug use (Buntain & Golbeck, 2015). Buntain & Golbeck described the architecture necessary to collect vast numbers of tweets over time based on a large number of search terms and the challenges that come with finding relevant information in the collected tweets. They then showed several examples of early analysis they have done on this data, examining temporal and geospatial trends. They also admitted that there were many challenges ahead in this space. Disambiguation is one of the largest. In addition, it was generally difficult to acquire recent and timely statistics on drug abuse across a wide geographic region.

As prescription drug abuse has become a major public health problem. Relationships and social context are important contributing factors. Social media provides online channels for people to build relationships that may influence attitudes and behaviors. To determine whether people who show signs of prescription drug abuse connect online with others who reinforce this behavior and to observe the conversation and engagement of these networks with regard to prescription drug abuse, Carl Lee Hanson studied exploration of social circles and prescription drug abuse through Twitter. He and his coworkers collected Twitter statuses mentioning prescription drugs from November 2011

to November 2012. From this set, 25 Twitter users were selected who discussed topics indicative of prescription drug abuse. Social circles of 100 people were discovered around each of these Twitter users; the tweets of the Twitter users in these networks were collected and analyzed according to prescription drug abuse discussion and interaction with other users about the topic. A strong correlation was found between the kinds of drugs mentioned by the index user and his or her network (mean $r=0.73$) and between the amount of interaction about prescription drugs and a level of abusiveness shown by the network ($r=0.85$, $P<.001$). Finally, they concluded that Twitter users who discuss prescription drug abuse online are surrounded by others who also discuss it—potentially reinforcing a negative behavior and social norm (Hanson, Cannon, Burton, & Giraud-Carrier, 2013).

Mental health is one of the essential parts of public health care. Glen Coppersmith, Mark Dredze, Craig Harman and Kristy Hollingshead examined a broad range of mental health conditions in Twitter data by identifying self-reported statements of diagnosis. They systematically explored language differences between ten conditions with respect to the general population, and to each other. Then they explored simple classifiers capable of distinguishing these users from their age- and gender-matched controls, based on signals quantified from the users' language. The classifiers also allowed them to systematically compare the language used by those with the ten conditions investigated, finding some groupings of the conditions found elsewhere in the literature, but not altogether obvious (Coppersmith, Dredze, Harman, & Hollingshead, 2015). They took this as evidence that examining mental health through the lens of language is fertile ground for advances in mental health writ large. The wealth of information encoded in continually-generated

social media is ripe for data scientists, computational linguists, and clinical psychologists, together, are well-positioned to drive this field forward.

Social media is a platform not only for individuals but also for government departments and organizations. Twitter can be used as a tool for educational organizations to inform the public about some issues like diabetes. Diabetes may affect one-third of US adults by 2050. Adopting a healthful diet and increasing physical activity are effective in preventing type 2 diabetes and decreasing the severity of diabetes-related complications. Educating and informing the public about health problems is a service provided by local health departments (LHDs). Jenine K. Harris et al. examined how LHDs are using social media to educate and inform the public about diabetes. In June 2012 they used NVivo 10 to collect all tweets ever posted from every LHD with a Twitter account and identified tweets about diabetes. They used a 2010 National Association of County and City Health Officials survey to compare characteristics of LHDs that tweeted about diabetes with those that did not. Content analysis was used to classify each tweet topic. They found that of 217 LHDs with Twitter accounts, 126 had ever tweeted about diabetes, with 3 diabetes tweets being the median since adopting Twitter. LHDs tweeting about diabetes were in jurisdictions with larger populations and had more staff and higher spending than LHDs not tweeting about diabetes. They were significantly more likely to employ a public information specialist and provide programs in diabetes-related areas. There was also a weak positive association between jurisdiction diabetes rate and the percentage of all tweets that were about diabetes ($r = .16$; $P = .049$). So they conclude LHDs were beginning to use social media to educate and inform their constituents about diabetes. An

understanding of the reach and effectiveness of social media could enable public health practitioners to use them more effectively (Harris, Mueller, Snider, & Haire-Joshu, 2013).

Public health-related tweets are difficult to identify in large conversational datasets like Twitter.com. Even more challenging is the visualization and analyses of the spatial patterns encoded in tweets. In Debarchana Ghosha and Rajarshi Guha's study of mapping tweets with topic modeling and geographic information system, obesity is chosen as a test theme to demonstrate the effectiveness of topic modeling using Latent Dirichlet Allocation (LDA) and spatial analysis using Geographic Information System (GIS) (Ghosh & Guha, 2013). The dataset was constructed from tweets (originating from the United States) extracted from Twitter.com on obesity-related queries. Examples of such queries are 'food deserts', 'fast food', and 'childhood obesity'. The tweets were also georeferenced and time-stamped. Three cohesive and meaningful themes such as 'childhood obesity and schools', 'obesity prevention', and 'obesity and food habits' are extracted from the LDA model. The GIS analysis of the extracted themes showed distinct spatial patterns between rural and urban areas, northern and southern states, and between coasts and inland states. Further, relating the themes with ancillary datasets such as the US census and locations of fast-food restaurants based upon the location of the tweets in a GIS environment opened new avenues for spatial analyses and mapping. Therefore the techniques used in this study provide a possible toolset for computational social scientists in general, and health researchers in specific, to better understand health problems from large conversational datasets.

In another study to explore the use of social media as a tool for health communication. Harris J.K. and his coworker used a mixed-methods design to examine communication about childhood obesity on Twitter (Harris et al., 2013). In their study,

NodeXL was used to collect tweets sent in June 2013 containing the hashtag #childhoodobesity. Tweets were coded for content; tweeters were classified by sector and health focus. Data were also collected on the network of follower connections among the tweeters. They used descriptive statistics and exponential random graph modeling to examine tweet content, characteristics of tweeters, and the composition and structure of the network of connections facilitating communication among tweeters. Eventually, they collected 1110 tweets originating from 576 unique Twitter users. More individuals (65.6%) than organizations (32.9%) tweeted. More tweets focused on individual behavior than environment or policy. Few governments and educational tweeters were in the network, but they were more likely than private individuals to be followed by others. After analyzing the results, they concluded that there was an opportunity to better disseminate evidence-based information to a broad audience through Twitter by increasing the presence of credible sources in the #childhoodobesity conversation and focusing the content of tweets on scientific evidence.

Another project created a Twitter classification model, which is aimed to design and test data collection and management tools that can be used to study the use of mobile fitness applications and social networking within the context of physical activity (Vickey, Ginis, & Dabrowski, 2013). That project was conducted over a 6-month period and involved collecting publically shared Twitter data from five mobile fitness apps (Nike+, RunKeeper, MyFitnessPal, Endomondo, and dailymile). During that time, over 2.8 million tweets were collected, processed, and categorized using an online tweet collection application and a customized JavaScript. Using the grounded theory, a classification model was developed to categorize and understand the types of information being shared by

application users. Their data showed that by tracking mobile fitness app hashtags, a wealth of information could be gathered to include but not limited to daily use patterns, exercise frequency, location-based workouts, and overall workout sentiment.

Those previous Twitter studies have dealt with extracting common topics of one health issue discussed by the users to better understand common themes. However, there is one study that utilized an innovative approach to computationally analyze unstructured health-related text data exchanged via Twitter to characterize health opinions regarding four common health issues, including diabetes, diet, exercise, and obesity (DDEO) on a population level (Karami et al., 2018). This study identified the characteristics of the most common health opinions with respect to DDEO and discloses public perception of the relationship between diabetes, diet, exercise, and obesity. These common public opinions/topics and perceptions can be used by providers and public health agencies to better understand the common opinions of their population denominators in regard to DDEO, and reflect upon those opinions accordingly.

Kathy Lee, Ankit Agrawal and Alok Choudhary's work about mining social media streams to improve public health allergy surveillance is the most inspiring paper for me. As mentioned in the paper, with the prevalence of social media, people sharing experiences and opinions on personal health symptoms and concerns on social media are increasing (Lee, Agrawal, & Choudhary, 2015). Mining those publicly available health-related data potentially can provide valuable healthcare insights. In this paper, the authors proposed a real-time allergy surveillance system that first classifies tweets to identify those that mention actual allergy incidents using the bag-of-words model and NaiveBayes Multinomial classifier and applies in-depth text and spatiotemporal analysis. They

collected allergy-related tweets from public tweet stream using twitter's streaming API. They had collected over 6.3 million tweets that mention 'allergy' or 'allergies' created by over 3.1 million unique users over 28 months from January 2013 to April 2015. And they used methods including data preprocessing, data classification, text mining, and Spatiotemporal Mining to get the result. Their experimental results showed that the proposed system can detect predominant allergy types with high precision and that allergy-related tweet volume is highly correlated to the weather data (daily maximum temperature).

In the past decade, with a dramatic increase in internet use, online data has been extensively used to retrieve health information and to detect disease activities. Web search queries data have been studied to track influenza activities. Ginsberg et al. used flu-related google search queries data to estimate current flu activity near real-time, 1-2 weeks in advance of the records by the traditional flu surveillance system. Recent research on public health and disease surveillance using online data has mostly focused on monitoring and predicting influenza levels. Researchers have used twitter data to monitor influenza outbreak and to predict flu activities. Lee et al. built a real-time disease surveillance system that uses Twitter data to track flu activity. Signorini et al. attempted estimating current influenza activity by tracking public sentiment and applying support vector machine algorithm on Twitter data generated during the Influenza A H1N1 pandemic. Chew et al. analyzed the content and sentiment of tweets generated during the 2009 H1N1 outbreak and showed the potential and feasibility of using social media to conduct infodemiology studies for public health. There are many others who have used Twitter data for flu outbreak detection. Unlike earlier researchers who used twitter for flu activity detection and

prediction, to the best of our knowledge, their work is the first attempt examining allergy activities using a large scale twitter stream.

Lee et al. classifies trending topics into 18 general categories using text-based and network-based models. Aramaki et al. proposed a Twitter-based influenza epidemics detection method that used Natural Language Processing (NLP) to filter out negative influenza tweets. Tuarob et al. used ensemble machine learning techniques to identify health-related messages in a heterogeneous pool of social media data. In this work, the authors used bag-of-words model and explored using four different machine learning algorithms to find the best model to classify tweets into those that mention actual allergy incidents and those that mention general awareness or information about allergy season.

In this paper, we focus on examining only allergy activity using a large Twitter stream collected over two years and show in-depth spatiotemporal analysis results. They also applied natural language processing techniques to automatically identify prevalent allergy types from Twitter contents.

This article is the first study that examines a large-scale social media stream for an in-depth analysis of allergy activities. And it provides many enlightening ideas for me.

Methods

Our approach uses statistical, semantic and linguistics analysis for disclosing health characteristics of opinions in tweets talking about allergy. The present study includes data collection, data preprocessing, topic discovery, and topic-content analysis.

Data Collection

Twitter Dataset

This phase collected tweets using Twitter's Application Programming Interfaces (API) (Twitter, 2017). Twitter's APIs provide both historic and real-time data collections. This paper adopted the historic method to collect publicly available English tweets from 01/01/2019 to 12/31/2019 using several pre-defined allergy-related queries. Within the Twitter API, allergy, hay fever, rhinitis, urticaria, anaphylaxis were selected as the related words and the related health areas. 2,189,597 unique tweets were collected by the query "allergy OR hay fever OR rhinitis OR urticaria OR anaphylaxis". Some talk about allergy types and symptoms (e.g. turns out I'm allergic to the new washing powder we bought & my body is covered in rashes and blisters). And there are others talking about their emotions and feelings negatively (e.g., this allergy really sucks) or positively (e.g., I love having allergic reactions to things on my face). Results for monthly distributions set out in Figure 1. Clearly, users of Twitter appear to post more tweets in the second quarter (666,513), particularly in April (251,581). Nonetheless, in February users tweeted less than in every other month.

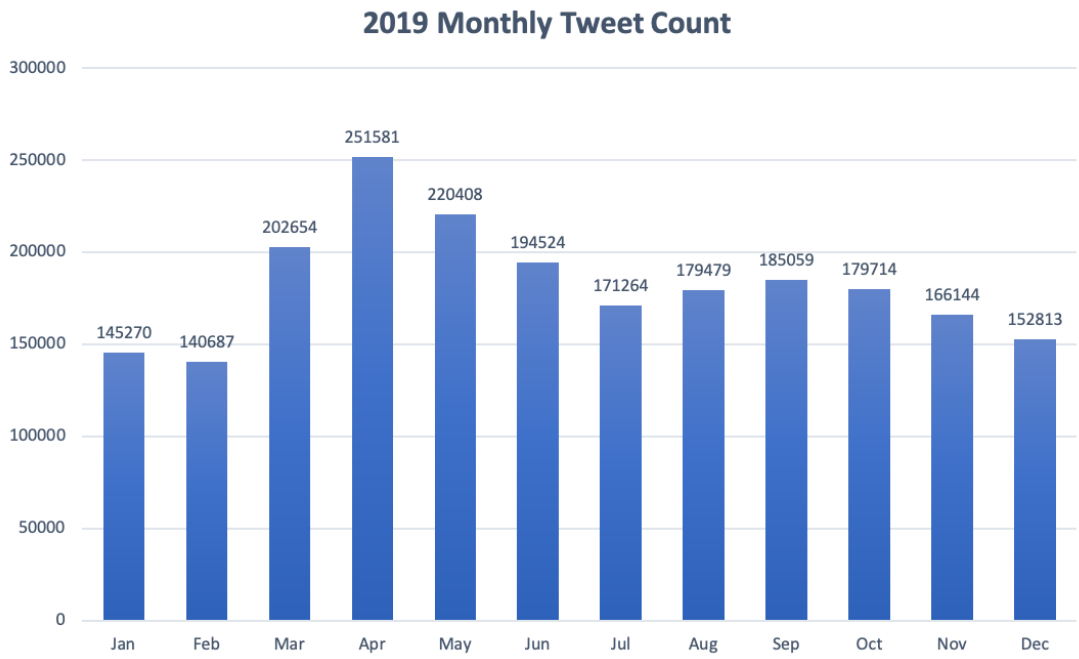

Figure 1: Allergy-related tweet count in 2019

Pollen Dataset

I collected monthly average pollen levels during 2019 for the US main cities from pollen.com. A pollen count is the measurement of the number of grains of pollen in a cubic meter of air. High pollen counts can sometimes lead to increased rates of an allergic reaction for those with allergic disorders. Usually, the counts are announced for specific plants such as grass, ash, or olive. These are tailored to common plants in the measured areas. Mild winters with warmer days lead to an increase in pollen counts while colder winters lead to delayed pollen release (Skinner, 2016). A pollen index defined by pollen.com is a number between 0 and 12 and divided into five categories: 0-2.4 (low), 2.5-4.8 (low-med), 4.9-7.2 (medium), 7.3-9.6 (med-high), 9.7-12.0 (high). The dataset contains 16 cities, which covers the top 10 cities ranking by population. The details can be seen in Table 1.

Table 1: Monthly pollen levels in the main cities of the US

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Chapel Hill	1.8	4.8	8.4	9.5	7.4	3.3	2.8	6.1	6.8	2.5	0.5	0.9
New York	0.1	1.3	5.6	9.3	7.7	5	4.6	8.7	4.8	1.2	0.2	0.1
Houston	5.2	7	9.5	7.2	5	3.5	2.7	4.3	7.1	5.3	2	4.8
Atlanta	1.8	4.8	9.7	8.5	6.1	3.2	3.2	6.3	8.6	3.1	1.2	1.5
Seattle	4.4	4.4	8.5	8	7.7	5.3	4.4	3.4	1.4	1.1	0.8	1.4
San Francisco	4.1	4	7	7.6	5.5	6.3	3.6	5.2	5.8	3.3	2	2.8
Denver	0.6	2	7.1	8.8	7.1	5.5	5.6	9.2	7.8	2.6	0.2	0.1
Los Angeles	7.1	6.4	7.9	7.3	4.2	4	3.2	5.3	6.9	4.9	2.5	3.2
Chicago	0.1	0.3	3.8	7.8	6.5	4	4.3	7.8	5.1	1.5	0.1	0.1
Phoenix	5.4	6.5	9.8	9.3	5.4	4.2	3.9	4.7	5.9	5.5	3.5	4
Philadelphia	0.1	1.5	6	9.3	7.4	4.4	4	8.4	4.9	1.1	0.1	0.1
San Antonio	8.4	7.8	9.4	8.2	5.1	4.4	3.6	7.2	9.6	6.2	3.5	8.7
San Diego	5.9	6.3	8.3	8.3	4.7	4.2	3.5	4.4	5.8	4.3	2.6	3.8
Dallas	7.5	7.8	9.3	8.2	5.5	3.7	4.2	6.6	10.4	6	2.2	8.4
San Jose	4.1	4	7	7.6	5.5	6.3	3.6	5.2	5.8	3.3	2	2.8
Dallas	7.5	7.8	9.3	8.2	5.5	3.7	4.2	6.6	10.4	6	2.2	8.4

Data Preprocessing

The preprocessing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process.

As I am interested in English messages, I have omitted tweets that are not written in English (7.6% of the initial data set, mostly Japanese). Emojis and punctuation are eliminated. All hyperlinks are replaced by the string ‘URL’ and all “@Username” are translated to "AT_USER". To further the data, I first lowercased all the text data.

Stopwords are the most common words in any natural language. Such stopwords may not bring any significance to the context of the document for the purpose of interpreting text data and building some NLP models. And in my study, stopwords are removed by using NLTK tools in Python.

Lemmatization was applied in order to reduce the size of the dictionary and thus the dimensionality of the description of text within the collection.

I also used n-gram to tokenize the tweets into consecutive sequences of words. To be specific, I added bigrams and trigrams to text data. And most notably, since the dataset is too large, I used the stratified random sampling method to extra 218,323 data with 2,812 unique tokens for topic modeling.

Topic Discovery

Topic modeling has a broad variety of applications in health and medical sciences such as forecasting protein-protein relationships based on the literature knowledge (Asou & Eguchi, 2008), finding applicable scientific principles and mechanisms in patients' health records (Arnold, El-Saden, Bui, & Taira, 2010), and identifying patterns of clinical events in a cohort of patients with brain cancer (Arnold & Speier, 2012).

Among all topic models, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is the most popular and most effective model (Lu, Mei, & Zhai, 2011; Paul & Dredze, 2011) as shown by studies that LDA is an effective computational linguistics model for exploration of topics in a corpus (Hong & Davison, 2010; Mcauliffe & Blei, 2008). LDA is defined as a generative probabilistic model for the collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics (Blei, Ng, & Jordan, 2003). It assumes a collection of K "topics." As seen in Figure 2, each topic describes a multinomial distribution over the vocabulary and is assumed to have been drawn from a Dirichlet β_k~Dirichlet(η). In the light of the topics, LDA assumes the following generative process for each document d. First, draw a distribution over topics

β_k~Dirichlet(α). Then, for each word i in the document, draw a topic index $z_{di} \in \{1, \ldots,$ K $\}$ from the topic weights $z_{di} \sim \theta_d$ and draw the observed word w_{di} from the selected topic, $w_{di} \sim \beta_{z_{di}}$.

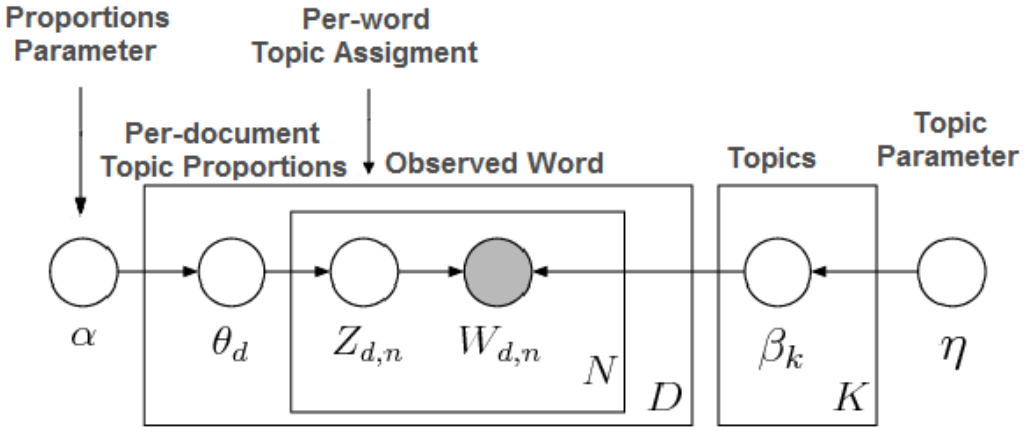

Figure 2: Latent Dirichlet Allocation (LDA) Structure, LDA represented as a graphical model in which the nodes denote the random variables and the edges of the dependencies between them. Unshaded nodes are unobserved or hidden variables and the shaded nodes represent the observed random variables. The boxes, called plates, indicate replication.

Twitter users can post their opinions or share information about a subject to the public. Identifying the main topics of users' tweets provides an interesting point of reference, but conceptualizing larger subtopics of millions of tweets can reveal valuable insight into users' opinions. To discover topics from the collected tweets, I set Gensim's standard LDA as the baseline model and compared the improved Mallet implementation of LDA (Blei et al., 2003; McCallum, 2002) with it. To determine the optimum number of topics, I used hyperparameter tuning. The best performance was determined 152 topics.

Content Analysis

The topic content analysis component used an objective interpretation approach with a lexicon-based approach to analyze the content of topics. The lexicon-based approach uses dictionaries to expose the semantic orientation of words on a topic. Linguistic Inquiry and Word Count (LIWC) is a linguistics research technique that reveals thoughts, feelings,

personality, and motivations in the corpus (Karami & Zhou, 2014a, 2014b, 2015). LIWC is a straightforward text analysis software that counts words in psychologically meaningful categories. Empirical results using LIWC demonstrate its potential to detect meaning in a wide variety of experimental contexts, including to show attentional focus, emotionality, social relationships, t thought patterns, and human variations (Tausczik & Pennebaker, 2010).

To interpret a text, out of the total number of words in the text, LIWC calculates the percentage of words in the text that match a dictionary word. Word frequency is calculated against a word count dictionary in terms of percentages, by using the following formula:

$$LIWC\ word\ frequency = \frac{word\ counts\ against\ a\ dictionary}{total\ word\ count\ in\ a\ text} \times 100\%$$

where a dictionary refers to a collection of words and word stems (sometimes even phrases) that represent or quantify particular linguistic features or psychological structures of research interest. A dictionary needs to be defined in advance. The number of words and word stems in a dictionary varies from several to multiple hundreds. For instance, the LIWC "affect" dictionary comprises 935 words and stems. LIWC offers up to 66 built-in dictionaries that represent specific text characteristics, ranging from linguistic processes to spoken categories (Pennebaker et al., 2007). Among all the built-in dictionaries in LIWC, the most relevant to public health research include social processes (e.g. family, friends, and humans), affective processes (e.g. positive emotion, negative emotion, anxiety, anger, and sadness), biological processes (e.g. body, health, sexual, and ingestion), and personal concerns (e.g. work, achievement, leisure, home, money, religion, and death) (Wang et al.,

2016). In this analysis, I mainly used biology and emotion dictionaries in LIWC to analyze the contents of allergy-related tweets.

Results

Trend Analysis

The average amount of tweets every day is 5,999 shown as the dotted line in Figure 3. The graph demonstrates the general allergy level trend over time. The allergy level is the highest in mid-April, declines in June and July, starts rising again in August, and reaches its local maximum point in September. Some other studies have observed similar seasonal variations (Lee, Agrawal, Choudhary, 2015).

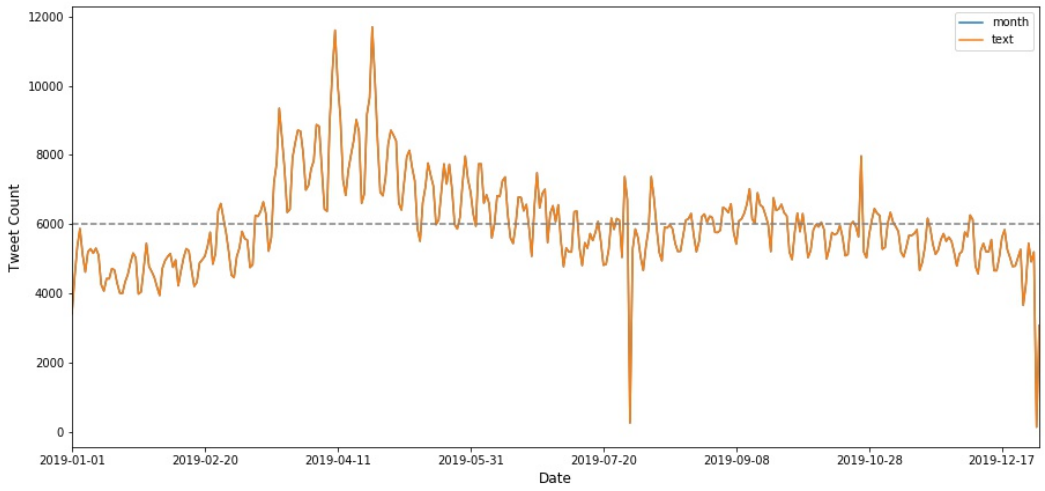

Figure 3: Daily data for allergy tweet count

As we know, pollen is one of the most common triggers of seasonal allergies. Many people know pollen allergy as “hay fever.” Experts typically refer to pollen allergy as “seasonal allergic rhinitis.” In this research, I compared monthly tweets count and pollen levels to see whether there is a correlation. Tweet count is strongly correlated to the pollen level with a correlation value of 0.699 (p-value<0.01). The correlation is clearly seen in

Figure 4. The two lines both climb from the beginning and reach their maximum points in April. Then the tweet count significantly declines as the pollen level is decreasing. They both have the local minimum points in July and keep rising to September. The trend also accords with common sense and some scientific explanations. It is understood that certain allergens, especially pollen, are seasonal. Tree pollen, for example, pops up in the spring (usually in late March to April), grass pollen occurs in the late spring (around May), weed pollen is most prevalent in the summer (July to August), and ragweed pollen takes over from summer to fall (late August to the first frost). This is consistent with the discernible patterns I found.

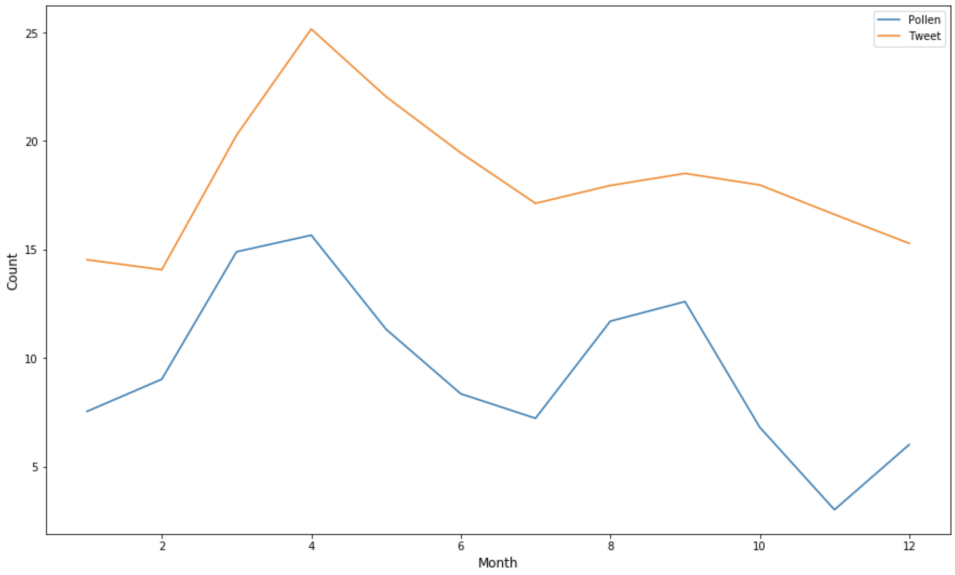

Figure 4: Monthly data/10000 for allergy tweet count (orange) and pollen level (blue) in 2019. The number of tweets is shrunk by a factor of 10000 for visualization.

And I also took Atlanta, GA as an example. When looking at the trend of the pollen count and tweet count during 2019, it is easy to find that the two lines follow identical trends which can be seen in Figure 5. Their peaks are both around April and they both have local maximum points in September. To determine the causal relationship, I ran a Granger causality test that is used to determine if one time series will be useful to forecast another.

The Null hypothesis is that the pollen level does not Granger-cause the Tweet amount. If the p-values are less than a significance level (0.05) then you reject the null hypothesis and conclude that the said lag of X is indeed useful. In my experiment, when I set the lag as 4, I got the smallest p-value, which is 0.0005. This indicates that the pollen amount affects tweet count and 4 lags of pollen should be included in this causality. The Pearson correlation coefficient between tweet count and the pollen level is 0.30 with p-value = 1.26e-08, which means that tweets count has a positive correlation with the pollen level.

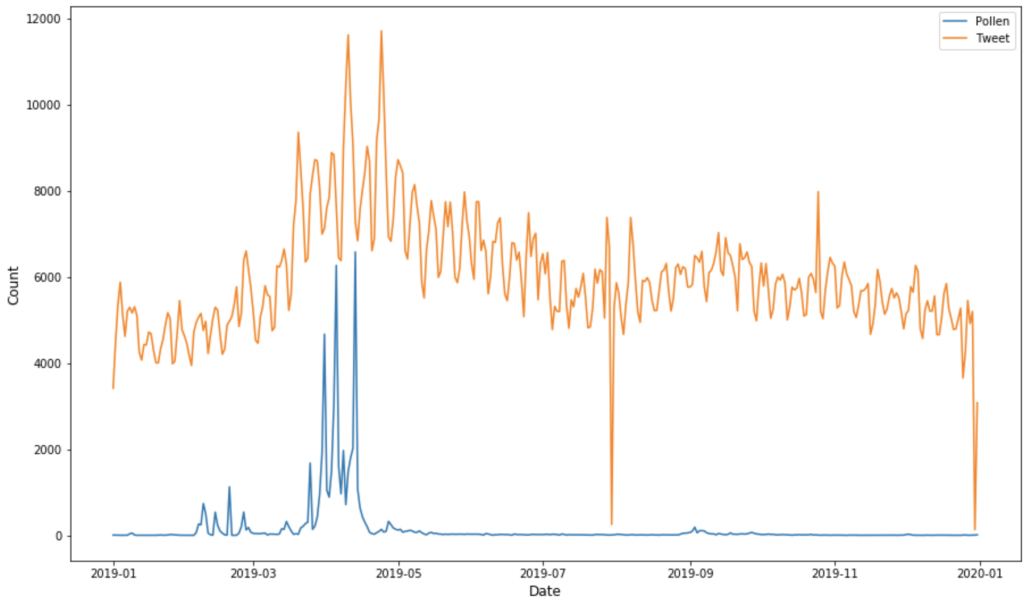

Figure 5: Atlanta pollen level and tweet count

Topic Modeling

Before identifying the opinions, the optimum number of topics needs to be found. I used GridSearch to determine the best hyperparameters. And the most important tuning parameter for LDA models is the number of topics. Out of all 218,323 allergy-related tweets returned by Tweeter's API, the highest log-likelihood from GridSearch was determined 425 topics.

Baseline

The first model I used is Gensim's ldamodel, which is also the baseline model I set. At 425 topics, Gensim had a coherence score of 0.48, a perplexity score of -15.34 . This is not great; indeed the Mallet algorithm which I looked at next almost always outperforms Gensim's. Mallet (MAchine Learning for LanguagE Toolkit) is a Java-based package put out by UMASS Amherst. The difference between Mallet and Gensim's standard LDA is that Gensim uses a Variational Bayes sampling method which is faster but less precise that Mallet's Gibbs Sampling.

To evaluate the results of Mallet version of LDA and compare it with Gensim's ldamodel, I used metrics like perplexity and coherence, along with eyeballing and model visualizations.

Eyeballing

The most common approach to assessing the quality of topics is the "eyeballing" approach, where topics are inspected deliberately and manually labelled (Morstatter and Liu, 2018).

The top 10 topics associated with its keywords and weights are displayed in Table 2. Taken the first topic as an example, it is a represented as 0.354*"side_effect" + 0.288*"side" + 0.258*"effect" + 0.022*"unfortunate" + 0.019*"milkshake" + 0.016*"painkiller" + 0.007*"people" + 0.006*"many" + 0.006*"allergie" + 0.006*"could". It means the top 10 keywords that contribute to this topic are: 'side_effect', 'unfortunate', 'milkshake', 'painkiller' and so on and the weight of 'side_effect' on this topic is 0.354. The weights reflect how important a keyword is to that topic. By manually distinguishing,

those topics can be interpreted as side effect, fruit, and other topics under the consideration of allergy.

Table 2: The top 10 topics generated from the LDA model.

Number	Representation	Label
1	0.354*"side_effect" + 0.288*"side" + 0.258*"effect" + 0.022*"unfortunate" + 0.019*"milkshake" + 0.016*"painkiller" + 0.007*"people" + 0.006*"many" + 0.006*"allergie" + 0.006*"could"	Side effect
2	0.186*"banana" + 0.119*"strawberry" + 0.110*"door" + 0.107*"dear" + 0.074*"piercing" + 0.061*"pickle" + 0.049*"brush" + 0.043*"mile" + 0.041*"donate" + 0.033*"refund"	Fruit
3	0.169*"daily" + 0.120*"heat" + 0.096*"expose" + 0.095*"price" + 0.091*"common_sense" + 0.064*"adhesive" + 0.061*"ban" + 0.055*"bump" + 0.041*"common" + 0.031*"sense"	Daily acts
4	0.170*"huge" + 0.155*"stop_sneeze" + 0.097*"rain" + 0.087*"dark" + 0.077*"painful" + 0.064*"stop" + 0.057*"sneeze" + 0.054*"pasta" + 0.054*"separate" + 0.035*"instantly"	Sneeze
5	0.368*"woman" + 0.088*"pregnant" + 0.082*"compare" + 0.063*"liquid" + 0.058*"memory" + 0.054*"suggestion" + 0.047*"loud" + 0.037*"washing" + 0.034*"alot" + 0.033*"outdoor"	Women
6	0.652*"allergic_reaction" + 0.272*"reaction" + 0.016*"cause" + 0.009*"give" + 0.008*"twitter" + 0.005*"go" + 0.004*"use" + 0.004*"time" + 0.004*"know" + 0.003*"lavender"	Allergy reaction
7	0.577*"take" + 0.126*"med" + 0.032*"onion" + 0.028*"help" + 0.023*"time" + 0.020*"almond" + 0.017*"need" + 0.017*"prescribe" + 0.016*"therapy" + 0.013*"ridiculous"	Medicine
8	0.310*"get" + 0.157*"sick" + 0.104*"stuff" + 0.056*"get_sick" + 0.036*"figure" + 0.032*"stupid" + 0.030*"put" + 0.029*"staff" + 0.027*"folk" + 0.026*"actually"	Sick
9	0.532*"food" + 0.120*"child" + 0.063*"people" + 0.049*"twitter" + 0.044*"many" + 0.039*"idea" + 0.018*"know" + 0.016*"give" + 0.016*"could" + 0.015*"foodallergie"	Food allergy
10	0.083*"think" + 0.081*"know" + 0.064*"go" + 0.063*"feel" + 0.063*"make" + 0.063*"really" + 0.054*"good" + 0.053*"want" + 0.050*"thing" + 0.046*"work"	Thoughts

Also, Figure 6 visualizes the first four of them are visualized using the world cloud approach. The word cloud visualizations are consistent with the keyword weights of each topic.

Figure 6: Word Clouds of Top 10 Keywords in Four Topics

I used pyLDAvis in Python to render a more picturesque and realistic visualization. It is the most widely used and a nice way to represent the information contained in a topic model. Figure 7 shows the visualization of the LDA model and the most salient terms. The left panel of the pyLDAvis graph presents a global view of the topic model. Within this perspective, the topics are plotted as circles in the two-dimensional plane whose centers are determined by computing the distance between topics, and then by using multidimensional scaling to project the intertopic distances onto two dimensions, as is done in (Chuang et al., 2012b). And the circle size is proportional to the topic's overall prevalence in the corpus.

In my model visualization, the topic circles are distributed evenly in the 2d plane with appropriate overlaps. The circle with the number 1 is the most prevalent one, accounting for 7.6% of the tokens. In the right panel, the saliency measure is used for ranking selecting relevant terms (Chuang et al., 2012a), which considers both term frequency and how much information the term can provide. The top 30 most salient terms

are allergic_reaction, food, take, reaction, year, would, tell, love, say, get, look, today, cat, much, never, come, peanut, life, thank, try, keep, skin, feel, severe, face, make, asthma, body, friend, eat. These words mainly describe allergens (e.g., food, cat, peanut), allergy symptoms (e.g., allergic_reaction, skin, asthma, face), and feelings (e.g., love, life, feel, severe). It covers almost all the topics when people are talking about allergies.

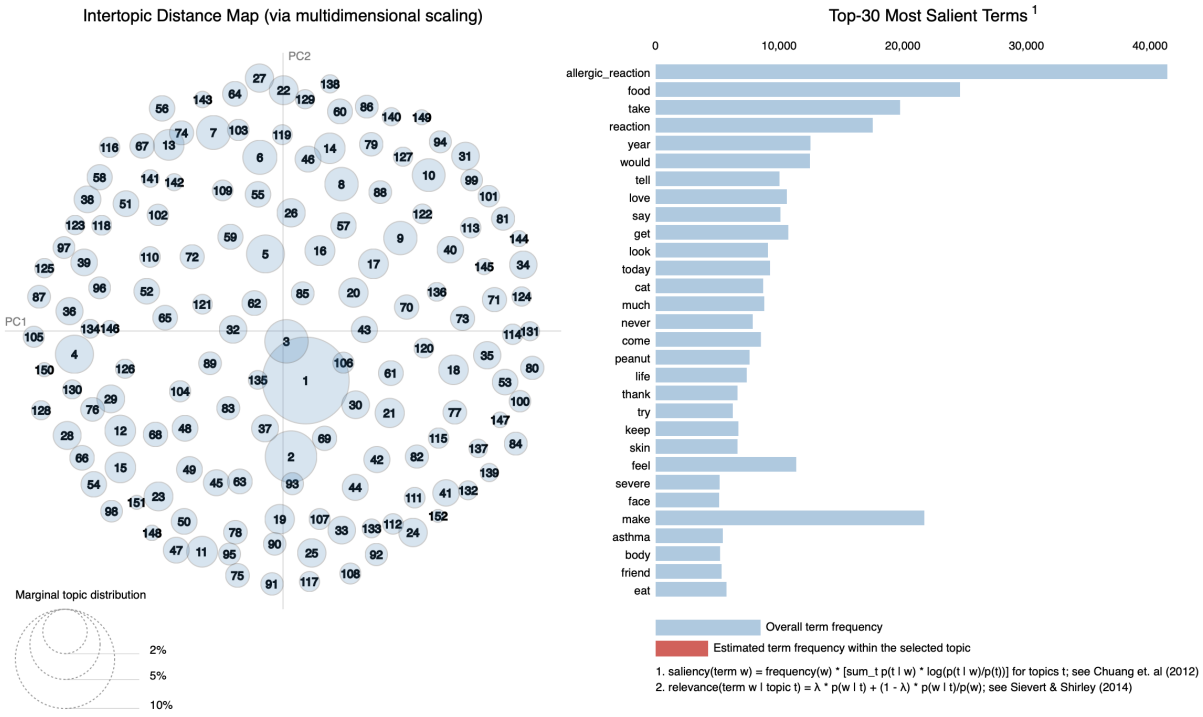

Figure 7: The layout of LDAvis, with the global topic view on the left, and the term bar charts on the right.

This method of evaluation, while common, has the issue that it is ad-hoc and time-consuming. It requires human labor to manually decide the quality of the results. This is a major problem in the topic assessment as this evaluation can be subjective, sometimes coming down to just one researcher who assigns definitions to the topics learned from the model. To mitigate this issue, researchers have investigated imposing principled measures for topic interpretability. Additionally, this has implications for reproducibility, as a

different researcher may have a different interpretation of the top words (Morstatter and Liu, 2018).

Topic Perplexity

Perplexity is one of the intrinsic evaluation metrics and is widely used for language model evaluation. It captures how surprised a model is of new data it has not seen before and is measured as the normalized log-likelihood of a held-out test set. Focussing on the log-likelihood part, you can think of the perplexity metric as measuring how probable some new unseen data is given the model that was learned earlier. In other words, how well does the model represent or reproduce the statistics of the held-out data. The lower the perplexity, the better the model.

The perplexity of my LDA model is -28.36, which means that the model is satisfactory from the perplexity point of view. However, however, Chang et al. (2009) found that perplexity does not always correlate with semantically interpretable topics. Predictive likelihood (or equivalently, perplexity) and human judgement are often not correlated, and even sometimes slightly anti-correlated. This limitation of perplexity measure acted as a catalyst a motivation for further research trying to model the human judgment, and hence topic coherence.

Topic Coherence

Topic coherence measures score a single topic by calculating the degree of semantic similarity between high scoring words in the topic. These metrics help differentiate between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. Among various coherence measures, I selected C_v implementation in my study. C_v measure is based on a sliding window, one-set segmentation of the top

words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity (Roder, Both and Hinneburg, 2015). The C_v coherence score is from 0 to 1, the higher, the better. I got 0.67 coherence in my model. From the perspective of the coherence score, the model proves itself again.

Content Analysis

This section will present and discuss the LIWC results in detail. The average word count of a tweet is 22.71 and the average number of words per sentence is13.40. Four summary variables from the LIWC analysis are used (Figure 8). The mean of "Analytical thinking" is 55.46, which is characterized by words suggesting logical, formal, or hierarchical thinking. Scores on the "Authenticity" summary variable (language that suggests revealing oneself in an honest way) is significantly high too, with the mean of 45.58. The next predominant dimension was "Authenticity" (mean=42.36), a variable that refers to confidence, leadership, or social status. According to the LIWC documentation, "a high number for Clout suggests that the author is speaking from the perspective of high expertise and is confident; low Clout numbers suggest a more tentative, humble, even anxious style" (Pennebaker et al., 2018). Scores on the "Emotional tone" (language suggesting either positive or negative emotion) were lower than other variables, which is 34.44.

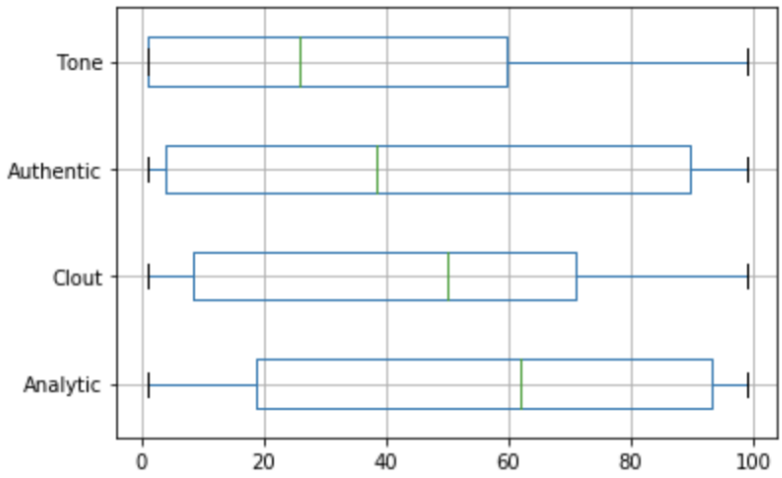

Figure 8: Summary variables of language style of allergy-related tweets (n=111899)

We also measured a number of other language dimensions, such as the use of words expressing different psychological processes, including emotional, social, perceptual and cognitive processes and relativity-related words. Among all those dimensions, sentiment analysis is an important part. The mean of positive emotion is 2.65 while the mean of negative emotion is 3.65. To determine if negative emotions overwhelm positives, a hypothesis test is necessary. The histograms in Figure 9 show the distribution of positive and negative emotions. Obviously, they are not normally distributed. If the data does not have the familiar Gaussian distribution, we must resort to the nonparametric version of the significance tests. These tests operate in a similar manner, but are distribution free, requiring that real-valued data be first transformed into rank data before the test can be performed. In this study, I assume the positive emotion scores and negative emotion scores are independent. In statistics, the Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from

a second population. After conducting a one-sided Mann–Whitney U test, the statistic is 6696975603.0000 and P<0.001. The p-value strongly suggests that the sample distributions are different, as is expected. As a result, it is true that LIWC's negative emotion was significantly higher than LIWC's positive emotion.

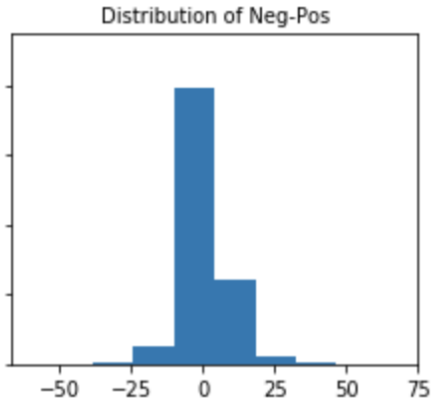

Figure 9: Histograms of positive and negative emotion scores

Moreover, there are some other interesting findings shown in Table 3. Allergy-related tweets have relatively high scores in biology and health aspects, which is very easy to understand. Also, people are talking about the present more, stating what they are thinking or feeling now.

Table 3: Biological process and time orientation parts of the LIWC result

Column1	Category	Mean	SD
Biological Processes	**bio**	8.439591	8.084892
	body	1.629591	3.710609
	health	5.991296	6.615233
	sexual	0.470662	2.546651
	ingest	0.495388	1.900555
Time Orientation	**focuspast**	2.629677	4.635333
	focuspresent	11.57177	8.673429
	focusfuture	0.993956	2.696392

Discussion

Allergy is a common belief relating to public health. The study of individual opinions by automated algorithmic techniques can be a useful method for better characterizing a population's health opinions. Traditional public health polls and polls are limited by small sample sizes; however, Twitter offers a forum for collecting a variety of views and exchanging information shared in the tweeter's language. Research indicates that there is a strong connection between Twitter health exchanges and the reports of the Centers for Disease Control and Prevention (Prier, Smith, Giraud- Carrier, & Hanson, 2011).

This research provides a computational content analysis approach to conduct in-depth analysis using a large data set of tweets. This framework decodes public health opinions in the case of allergy-related tweets that can be applied to other public health issues.

The time-series analysis showed the relationship between tweets and pollen. They are highly correlated. The pollen amount affects tweet count and 4 lags of pollen should be included in this causality. Also, I tracked how the trend of mentions of two different allergy types differ over time. The tweet volume mentioning ‘pollen allergy’ rises very high during the spring and the fall and remains very low in the summer. However, unlike pollen allergy, the tweet volume mentioning ‘peanut allergy’ stays relatively constant throughout the year. This observation implies that the seasonality observed in overall allergy dataset in figure 4 and 5. What’s more, further research like allergy surveillance can be conducted based on

this finding. This paper addresses a need for clinical providers, public health experts, and social scientists to utilize a large conversational data set to collect and utilize population-level opinions and information needs. Although our framework is applied to Twitter, the applications from this study can be used in patient communication devices monitored by allergists with social media accounts, and support large scale population-wide initiatives to help prevent allergy and alleviate allergic symptoms.

This research has some limitations. First, this study has some problems when collecting data. It does not include nationwide pollen statistics. Certificated pollen counting stations are located in every state and every county. It is difficult to collect the historical pollen data of 2019 from all the stations and there is no universal definition of pollen count at the nation-level. What it did is collecting 16 presentative cites and calculating their mean pollen count as the national level while pollen levels may vary due to regional differences. This analysis does not take the geographical location of Twitter users into consideration either. Thus it does not reveal if certain geographical differences exist. Also, I used a limited number of queries to select the initial pool of tweets, thus perhaps missing tweets that may have been relevant to allergy but have used unusual terms referenced. And the analysis only included tweets generated in 2019; however, public opinion can change during years. Additionally, we did not track individuals across time to detect changes in common themes discussed.

Second, this study does not evaluate the LDA model very thoroughly. LDA is popular for text analysis, providing both a predictive and latent topic representation of the corpus. However, there is a longstanding assumption that the latent space discovered by these models is generally meaningful and useful and that evaluating such assumptions is

challenging due to its unsupervised training process. Besides, there is a no-gold standard list of topics to compare against every corpus. Nevertheless, it is equally important to identify if a trained model is objectively good or bad, as well have an ability to compare different models/methods. To do so, one would require an objective measure for quality. Traditionally, and still for many practical applications, to evaluate if "the correct thing" has been learned about the corpus, implicit knowledge and "eyeballing" approaches are used. But the "traditional" approached are too subjective and inefficient. Moreover, the perplexity metric has the limitation that it is not strongly correlated to human judgment and even sometimes slightly anti-correlated (Chang et al., 2009). Chang et al. ran a large scale experiment on the Amazon Mechanical Turk platform (2009). They ran a large scale experiment on the Amazon Turk platform. For each topic, they took the five top words of those topics and added a random sixth word. Then, they presented these lists of six words to people asking them which is the intruder word. If all the people asked could tell which is the intruder, then we can conclude safely that the topic is good at describing an idea. If on the other hand, many people identified other words as the intruder, it means that they could not see the logic into the association of words, and we can conclude the topic was not good enough. The result proves that, given a topic, the five words that have the largest frequency withing their topic are usually not good at describing one coherent idea; at least not good enough to be able to recognize an intruder. And, optimizing for perplexity may not yield human interpretable topics. As to the coherence score, it is the most advanced and appropriate measure. This study got a 0.67 coherence score, and the results would be better if 0.7 or higher can be reached.

Third, this study took an unsupervised approach to analyze the content of the tweets, especially sentiments using LIWC. And some researches show that many other tools have better performance than the LIWC tool. For example, Crossley et al. tested a new sentiment analysis tool, SEANCE, which is freely available to researchers and provides an automated approach to the examination of discourse in terms of sentiment, cognition, and social order, against the most common tool used in sentiment analysis for behavioral studies (LIWC) and found that both the individual indices and the component scores statistically outperformed LIWC in classic sentiment analysis tasks. What's more, LIWC provides only positive and negative emotions. Nevertheless, sentiments could also be extremely positive/negative, slightly positive/negative or neutral. Last but not least, lexicon-based sentiment analysis has its own limitations, like no training and no learning capacity.

Conclusion and Future Work

People eagerly share and post content on social media expressing their points of view in an unrestricted way. The dimensionality and size of opinionated data are growing exponentially and turn out to be valuable sources for text mining. This study represents one of the early steps in developing routine processes to collect, analyze, and interpret allergy-related posts to social media around health-related topics and presents a transdisciplinary approach to analyzing public discussions around health topics.

This paper discovered the trends of allergy-related tweets using time series analysis, clustered the topics by the LDA model and analyzed the contents via the LIWC2015 tool in order to answer the research questions in the beginning. Taken the monthly pollen count as ground truth data, the Pearson correlation coefficient between pollen level and tweets count is 0.699 (p-value<0.01), which means there is a significant correlation between these two attributes. The Granger causality test reveals that pollen counts change cause the tweets count to vary. It's worth noting that pollen count and tweets count both have seasonal trends, which might be explained by the weather/temperature. In the future, I would like to conduct a time-series analysis of the daily temperature to find if there is any correlation between temperature and tweet count. To solve the second research question, I built an LDA model to cluster the allergy-related tweets into 152 topics. And I used three approached to evaluate the model, eyeballing, perplexity (-28.36) and coherence (0.67). As the limitation of the computation power, this research is based on only 10% of the data I crawled. However, the result is quite convincing even with the small size. Applying to a larger dataset will more

likely achieve better results. The LIWC2015 tool is used to analyze the contents of tweets. Especially, sentiment analysis is conducted to learn about the emotions of twitter users when talking about allergies. And it turns out that people express negative emotions than positives ones. Sentiment analysis using the LIWC is essentially lexicon-based. And I plan to manually label the data and utilize machine learning techniques to do a supervised classification job.

With billions of social media users, the ability to collect and synthesize social media data will continue to grow. Considering that, my future research plans include introducing a dynamic framework to collect and analyze allergy-related tweets during extended time periods (multiple years) and incorporating spatial analysis of allergy-related tweets. And developing methods to make this process more streamlined and robust will allow for more rapid identification of public health trends in real-time.

References

American College of Allergy, Asthma, and Immunology. *Allergy Facts.* http://acaai.org/news/facts-statistics/allergies (Retrieved March 14 2018)

A. Choudhary, W. Hendrix, K. Lee, D. Palsetia, and W.-K. Liao. Social media evolution of the egyptian revolution. *Commun. ACM, 55*(5):74–80, May 2012.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces (pp. 74-77).

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., et al. (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological Science, 26(2), 159–169.

Gohil, S., Vuik, S., & Darzi, A. (2018). Sentiment analysis of health care tweets: review of the methods used. JMIR public health and surveillance, 4(2), e43.

Jason Chuang, Daniel Ramage, Christopher D. Manning and Jeffrey Heer. 2012a. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. CHI.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1 – 8, 2011.

Kapadia, S. (2019, August 19). Evaluate Topic Models: Latent Dirichlet Allocation (LDA). Retrieved from https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H. H. K., & Shaw, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on Twitter. International Journal of Information Management, 38(1), 1-6. https://doi.org/10.1016/j.ijinfomgt.2017.08.002

Lee, K., Agrawal, A., & Choudhary, A. (2015, August). Mining social media streams to improve public health allergy surveillance. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (pp. 815-822).

McConnell, T. H. (2013). *The nature of disease: pathology for the health professions.* Lippincott Williams & Wilkins.

Mejova, Y., Weber, I., & Macy, M. W. (2015). Twitter: A digital socioscope. Cambridge University Press.

Morstatter, F., & Liu, H. (2017). In search of coherence and consensus: measuring the interpretability of statistical topics. The Journal of Machine Learning Research, 18(1), 6177-6208.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on knowledge capture. ACM70–77.

Paul, M. J., & Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. ICWSM, Vol. 20265–272.

Paul, M. J., & Dredze, M. (2012). A model for mining public health topics from Twitter. Health, 11 16-6.

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.

Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahwah, NJ: Lawrence Erlbaum Associates.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The Development and Psychometric Properties of LIWC 2015. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf. Accessed 22 Mar 2018.

Pleplé, Q. (2013, May). Perplexity To Evaluate Topic Models. Retrieved from http://qpleple.com/perplexity-to-evaluate-topic-models/

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health-related topics on Twitter. International conference on social computing, behavioral-cultural modeling, and prediction. Springer18–25.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

Skinner, A. (2016, August 6). What is Pollen? Master Gardener Newspaper, p. D5.

Statista (2019). Number of social media users worldwide from 2010 to 2020. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Summary Health Statistics Tables for U.S. Adults: National Health Interview Survey, 2017, Tables A-2b, A-2c https://robynobrien.com/25-billion-cost-food-allergies/

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. ICWSM 10.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In In Proceedings of the Nineteenth International WWW Conference (WWW2010). ACM, 2010.

Wang, W., Hernandez, I., Newman, D. A., He, J., & Bian, J. (2016). Twitter analysis: Studying US weekly trends in work stress and emotion. Applied Psychology, 65(2), 355-378.

World Health Organization. White Book on Allergy 2011-2012 Executive Summary. By Prof. Ruby Pawankar, MD, PhD, Prof. Giorgio Walkter Canonica, MD, Prof. Stephen T. Holgate, BSc, MD, DSc, FMed Sci and Prof. Richard F. Lockey, MD.

Appendix

Table 1: Allergy-related Tweets count summary

Quarter	Quarter Summary	Month	Month Summary
Q1	488,611	**January**	145,270
		February	140,687
		March	202,654
Q2	666,513	**April**	251,581
		May	220,408
		June	194,524
Q3	535,802	**July**	171,264
		August	179,479
		September	185,059
Q4	498,671	**October**	179,714
		November	166,144
		December	152,813

Table 2 All topics from LDA modeling

Number	Representation
1	0.170*"huge" + 0.155*"stop_sneeze" + 0.097*"rain" + 0.087*"dark" + 0.077*"painful" + 0.064*"stop" + 0.057*"sneeze" + 0.054*"pasta" + 0.054*"separate" + 0.035*"instantly"
2	0.246*"check" + 0.241*"suffer" + 0.149*"room" + 0.054*"rise" + 0.043*"everyday" + 0.039*"twitter" + 0.031*"full_blow" + 0.029*"plate" + 0.020*"auto" + 0.017*"antibody"
3	0.134*"swell" + 0.129*"happy" + 0.128*"definitely" + 0.114*"seafood" + 0.084*"lot" + 0.055*"sit" + 0.049*"eye_swell" + 0.046*"begin" + 0.034*"flavor" + 0.032*"irritate"
4	0.469*"love" + 0.085*"one" + 0.084*"restaurant" + 0.055*"young" + 0.048*"special" + 0.046*"pizza" + 0.038*"dish" + 0.035*"prescription" + 0.027*"twitter" + 0.021*"limit"
5	0.310*"get" + 0.157*"sick" + 0.104*"stuff" + 0.056*"get_sick" + 0.036*"figure" + 0.032*"stupid" + 0.030*"put" + 0.029*"staff" + 0.027*"folk" + 0.026*"actually"

6	0.346*"friend" + 0.092*"important" + 0.060*"prevent" + 0.049*"bother" + 0.047*"sting" + 0.044*"cake" + 0.041*"remind" + 0.035*"effective" + 0.031*"surprise" + 0.027*"cow_milk"
7	0.173*"enjoy" + 0.152*"pass" + 0.101*"couple" + 0.084*"pack" + 0.081*"careful" + 0.076*"wine" + 0.062*"truly" + 0.046*"massive" + 0.030*"bathroom" + 0.030*"cinnamon"
8	0.267*"bad" + 0.149*"must" + 0.133*"disease" + 0.077*"explain" + 0.044*"color" + 0.031*"extreme" + 0.028*"get_bad" + 0.028*"ball" + 0.024*"related" + 0.024*"make"
9	0.177*"pick" + 0.120*"spot" + 0.112*"assume" + 0.100*"obviously" + 0.091*"wonderful" + 0.066*"properly" + 0.058*"soup" + 0.044*"inflammatory" + 0.039*"itchy_watery" + 0.033*"ticket"
10	0.368*"woman" + 0.088*"pregnant" + 0.082*"compare" + 0.063*"liquid" + 0.058*"memory" + 0.054*"suggestion" + 0.047*"loud" + 0.037*"washing" + 0.034*"alot" + 0.033*"outdoor"
11	0.193*"apparently" + 0.136*"plan" + 0.100*"run" + 0.061*"glass" + 0.059*"warn" + 0.035*"deodorant" + 0.032*"comfortable" + 0.028*"sleepy" + 0.026*"bright" + 0.026*"chinese"
12	0.286*"medication" + 0.161*"finally" + 0.103*"travel" + 0.072*"panic_attack" + 0.056*"inhaler" + 0.051*"panic" + 0.044*"breakfast" + 0.039*"recall" + 0.034*"chip" + 0.028*"attack"
13	0.197*"literally" + 0.166*"can" + 0.123*"anaphylaxi" + 0.120*"fine" + 0.068*"course" + 0.057*"react" + 0.045*"mostly" + 0.027*"science" + 0.020*"training" + 0.020*"standard"
14	0.209*"season" + 0.109*"walk" + 0.083*"affect" + 0.072*"kick" + 0.069*"anyway" + 0.055*"quality" + 0.050*"raise" + 0.042*"prefer" + 0.040*"flare" + 0.036*"environment"
15	0.125*"listen" + 0.118*"funny" + 0.115*"joke" + 0.088*"actual" + 0.084*"gift" + 0.062*"excuse" + 0.046*"music" + 0.034*"people" + 0.030*"plain" + 0.029*"active"
16	0.187*"itchy" + 0.176*"fever" + 0.167*"super" + 0.107*"absolutely" + 0.091*"save" + 0.047*"save_life" + 0.045*"itchy_eye" + 0.038*"beer" + 0.030*"anyways" + 0.024*"seizure"
17	0.250*"probably" + 0.243*"turn" + 0.127*"cook" + 0.078*"news" + 0.064*"hopefully" + 0.034*"ride" + 0.024*"final" + 0.022*"time" + 0.020*"addict" + 0.015*"official"
18	0.180*"less" + 0.174*"grow" + 0.097*"alcohol" + 0.088*"white" + 0.049*"learn_twitt" + 0.046*"current" + 0.037*"impossible" + 0.036*"people" + 0.034*"weed" + 0.032*"make"
19	0.253*"test" + 0.142*"already" + 0.118*"believe" + 0.072*"money" + 0.056*"relate" + 0.053*"local" + 0.048*"fear" + 0.046*"honey" + 0.044*"local_honey" + 0.028*"positive"
20	0.149*"symptom" + 0.136*"learn" + 0.131*"treat" + 0.114*"include" + 0.098*"common" + 0.057*"twitter" + 0.046*"report" + 0.029*"help" + 0.029*"horse" + 0.024*"bomb"
21	0.106*"safe" + 0.091*"able" + 0.077*"egg" + 0.076*"close" + 0.073*"wonder" + 0.064*"damn" + 0.041*"always" + 0.040*"choose" + 0.035*"dinner" + 0.035*"stand"

22	0.242*"trigger" + 0.126*"write" + 0.069*"large" + 0.062*"accept" + 0.058*"potential" + 0.052*"cut" + 0.048*"decision" + 0.045*"shed" + 0.044*"exercise" + 0.036*"destroy"
23	0.299*"eat" + 0.281*"stop" + 0.065*"worry" + 0.060*"stop_eat" + 0.044*"nee" + 0.028*"count" + 0.027*"evidence" + 0.021*"condom" + 0.015*"personality" + 0.012*"flat"
24	0.236*"understand" + 0.162*"study" + 0.107*"mouth" + 0.073*"owner" + 0.053*"people" + 0.041*"host" + 0.038*"crab" + 0.033*"eventually" + 0.031*"blind" + 0.029*"stage"
25	0.198*"kill" + 0.155*"hair" + 0.133*"shot" + 0.080*"honestly" + 0.076*"cross_contamination" + 0.052*"chance" + 0.044*"cross" + 0.034*"steroid" + 0.031*"contamination" + 0.029*"could"
26	0.243*"month" + 0.130*"part" + 0.120*"usually" + 0.065*"time" + 0.061*"anaphylaxis" + 0.055*"last" + 0.036*"last_time" + 0.035*"injection" + 0.020*"shirt" + 0.019*"teal"
27	0.228*"sure" + 0.197*"make_sure" + 0.172*"problem" + 0.100*"pretty" + 0.094*"pretty_sure" + 0.085*"make" + 0.015*"chef" + 0.015*"flea" + 0.014*"curious" + 0.012*"land"
28	0.157*"easy" + 0.151*"healthy" + 0.102*"recipe" + 0.093*"sugar" + 0.081*"disorder" + 0.056*"medical_condition" + 0.048*"helpful" + 0.046*"autism" + 0.041*"artificial" + 0.039*"account"
29	0.126*"drop" + 0.123*"clear" + 0.085*"metal" + 0.079*"unfortunately" + 0.073*"sauce" + 0.057*"nickel" + 0.056*"earring" + 0.035*"wear" + 0.033*"pierce" + 0.033*"silver"
30	0.247*"health" + 0.155*"treatment" + 0.094*"recently" + 0.063*"twitter" + 0.062*"testing" + 0.047*"improve" + 0.044*"update" + 0.032*"release" + 0.032*"trial" + 0.032*"infant"
31	0.125*"cream" + 0.116*"increase" + 0.084*"similar" + 0.072*"build" + 0.067*"movie" + 0.063*"future" + 0.059*"finger" + 0.046*"increase_risk" + 0.044*"tend" + 0.038*"reality"
32	0.319*"ever" + 0.192*"send" + 0.131*"chicken" + 0.093*"sign" + 0.039*"beef" + 0.029*"irritation" + 0.028*"relationship" + 0.024*"message" + 0.016*"coughing" + 0.015*"grill"
33	0.189*"risk" + 0.137*"lactose_intolerant" + 0.077*"intolerant" + 0.075*"reduce" + 0.072*"lactose" + 0.060*"around" + 0.057*"exposure" + 0.056*"sufferer" + 0.030*"people" + 0.028*"project"
34	0.287*"case" + 0.199*"taste" + 0.122*"provide" + 0.056*"ring" + 0.055*"personal" + 0.043*"request" + 0.026*"swallow" + 0.024*"judge" + 0.024*"fix" + 0.023*"possibility"
35	0.168*"serious" + 0.128*"word" + 0.087*"dietary_restriction" + 0.084*"force" + 0.062*"dietary" + 0.060*"rare" + 0.053*"restriction" + 0.051*"spread" + 0.044*"awareness" + 0.044*"people"
36	0.366*"nose" + 0.120*"runny_nose" + 0.083*"headache" + 0.073*"scare" + 0.060*"runny" + 0.043*"birth" + 0.038*"birth_control" + 0.023*"routine" + 0.021*"difficulty" + 0.019*"salmon"
37	0.207*"sorry" + 0.121*"brand" + 0.093*"awful" + 0.047*"let" + 0.047*"realise" + 0.040*"tape" + 0.038*"kitten" + 0.029*"awake" + 0.026*"fatigue" + 0.024*"wide"

38	0.267*"care" + 0.114*"agree" + 0.107*"rather" + 0.065*"emergency" + 0.060*"immediately" + 0.058*"urgent_care" + 0.044*"opinion" + 0.039*"appear" + 0.032*"knowledge" + 0.028*"urgent"
39	0.176*"option" + 0.135*"result" + 0.091*"information" + 0.076*"item" + 0.058*"solution" + 0.040*"twitter" + 0.029*"town" + 0.029*"sale" + 0.025*"lay" + 0.025*"suffer_seasonal"
40	0.463*"happen" + 0.090*"whenever" + 0.055*"review" + 0.053*"expert" + 0.047*"boy" + 0.037*"shake" + 0.035*"thing" + 0.035*"target" + 0.025*"mate" + 0.024*"match"
41	0.517*"try" + 0.071*"company" + 0.061*"soap" + 0.054*"laugh" + 0.042*"afford" + 0.042*"fragrance" + 0.025*"min" + 0.021*"make" + 0.017*"caffeine" + 0.017*"goodness"
42	0.336*"asthma" + 0.105*"often" + 0.072*"dust_mite" + 0.034*"mite" + 0.034*"dust" + 0.032*"pray" + 0.029*"bless" + 0.028*"twitter" + 0.027*"complete" + 0.025*"cousin"
43	0.214*"school" + 0.195*"deathly_allergic" + 0.089*"mention" + 0.069*"deathly" + 0.067*"carry" + 0.065*"try_kill" + 0.055*"high_school" + 0.035*"teacher" + 0.032*"hide" + 0.025*"jump"
44	0.258*"last_night" + 0.214*"night" + 0.172*"last" + 0.063*"basically" + 0.058*"horrible" + 0.047*"husband" + 0.033*"annoying" + 0.023*"survive" + 0.019*"heavy" + 0.018*"arrive"
45	0.626*"year" + 0.132*"last_year" + 0.052*"last" + 0.042*"time" + 0.029*"lack" + 0.026*"finish" + 0.022*"fry" + 0.021*"specialist" + 0.009*"habit" + 0.008*"still"
46	0.316*"sneeze" + 0.124*"death" + 0.106*"crazy" + 0.087*"stress" + 0.044*"floor" + 0.039*"somehow" + 0.037*"lately" + 0.028*"nose_run" + 0.020*"stuffy_nose" + 0.017*"direct"
47	0.103*"s" + 0.096*"strong" + 0.092*"brother" + 0.083*"cure" + 0.080*"quite" + 0.068*"claim" + 0.066*"next_week" + 0.062*"picture" + 0.058*"old" + 0.056*"feeling"
48	0.262*"change" + 0.173*"next" + 0.083*"hold" + 0.076*"epipen" + 0.070*"office" + 0.063*"notice" + 0.045*"time" + 0.030*"pink" + 0.021*"carry_epipen" + 0.020*"determine"
49	0.456*"peanut" + 0.133*"peanut_butter" + 0.111*"butter" + 0.096*"pet" + 0.080*"allergen" + 0.034*"people" + 0.024*"mildly" + 0.020*"mildly_allergic" + 0.011*"slice" + 0.007*"make"
50	0.228*"enough" + 0.132*"natural" + 0.075*"relief" + 0.062*"source" + 0.059*"past" + 0.046*"reply" + 0.045*"rescue" + 0.040*"past_day" + 0.036*"grade" + 0.036*"calm"
51	0.354*"side_effect" + 0.288*"side" + 0.258*"effect" + 0.022*"unfortunate" + 0.019*"milkshake" + 0.016*"painkiller" + 0.007*"people" + 0.006*"many" + 0.006*"allergie" + 0.006*"could"
52	0.154*"become" + 0.094*"fruit" + 0.090*"contain" + 0.070*"amazing" + 0.062*"sister" + 0.051*"struggle" + 0.046*"bear" + 0.039*"vegan" + 0.035*"bake" + 0.035*"mushroom"
53	0.157*"head" + 0.141*"wrong" + 0.137*"medical" + 0.084*"mess" + 0.070*"lead" + 0.067*"last_day" + 0.059*"advice" + 0.044*"salt" + 0.036*"pork" + 0.027*"last"

54	0.189*"protein" + 0.132*"rest" + 0.122*"level" + 0.087*"big" + 0.085*"milk_protein" + 0.066*"safety" + 0.035*"cater" + 0.034*"researcher" + 0.034*"function" + 0.030*"suit"
55	0.177*"type" + 0.081*"control" + 0.066*"difference" + 0.058*"simple" + 0.053*"expect" + 0.049*"dander" + 0.048*"people" + 0.046*"tongue" + 0.043*"easily" + 0.043*"ignore"
56	0.172*"anaphylactic_shock" + 0.166*"cover" + 0.113*"anaphylactic" + 0.109*"shock" + 0.064*"pull" + 0.042*"environmental" + 0.032*"style" + 0.025*"garden" + 0.024*"contribute" + 0.022*"surround"
57	0.230*"wait" + 0.167*"minute" + 0.135*"shellfish" + 0.104*"mine" + 0.061*"hypoallergenic" + 0.058*"upset" + 0.044*"kiss" + 0.033*"angry" + 0.027*"particularly" + 0.019*"band"
58	0.105*"end" + 0.091*"extremely" + 0.089*"good_luck" + 0.065*"short" + 0.062*"foot" + 0.058*"luck" + 0.055*"bottle" + 0.054*"situation" + 0.050*"appreciate" + 0.046*"trust"
59	0.239*"kid" + 0.129*"parent" + 0.057*"family_member" + 0.054*"though" + 0.053*"drive" + 0.048*"member" + 0.048*"diagnose" + 0.044*"process" + 0.044*"response" + 0.033*"trip"
60	0.187*"cute" + 0.121*"swear" + 0.065*"slight" + 0.056*"publish" + 0.054*"particular" + 0.051*"female" + 0.042*"inject" + 0.042*"purchase" + 0.041*"peel" + 0.038*"twitter"
61	0.215*"immune_system" + 0.200*"baby" + 0.141*"system" + 0.114*"immune" + 0.061*"form" + 0.044*"present" + 0.041*"tomato" + 0.029*"bacteria" + 0.023*"professional" + 0.018*"healthcare"
62	0.234*"ask" + 0.137*"question" + 0.102*"early" + 0.081*"service" + 0.078*"shrimp" + 0.058*"answer" + 0.041*"potato" + 0.025*"access" + 0.022*"track" + 0.022*"customer_service"
63	0.186*"banana" + 0.119*"strawberry" + 0.110*"door" + 0.107*"dear" + 0.074*"piercing" + 0.061*"pickle" + 0.049*"brush" + 0.043*"mile" + 0.041*"donate" + 0.033*"refund"
64	0.181*"sound" + 0.176*"patient" + 0.092*"label" + 0.087*"adult" + 0.086*"pineapple" + 0.035*"small_amount" + 0.035*"sniffle" + 0.031*"scream" + 0.030*"american" + 0.029*"spice"
65	0.495*"severe" + 0.148*"meal" + 0.068*"poison" + 0.059*"twice" + 0.047*"partner" + 0.038*"detail" + 0.023*"hotel" + 0.023*"balloon" + 0.019*"announce" + 0.016*"could"
66	0.357*"hand" + 0.155*"wash" + 0.093*"wash_hand" + 0.054*"everytime" + 0.051*"convince" + 0.045*"piss" + 0.045*"fave" + 0.031*"snicker" + 0.029*"lesson" + 0.025*"circle"
67	0.083*"think" + 0.081*"know" + 0.064*"go" + 0.063*"feel" + 0.063*"make" + 0.063*"really" + 0.054*"good" + 0.053*"want" + 0.050*"thing" + 0.046*"work"
68	0.526*"look" + 0.061*"benadryl" + 0.060*"lip" + 0.057*"entire" + 0.054*"take_benadryl" + 0.034*"paper" + 0.018*"addition" + 0.018*"make" + 0.017*"bruh" + 0.016*"abuse"
69	0.124*"break" + 0.109*"hive" + 0.108*"morning" + 0.100*"take_care" + 0.073*"break_hive" + 0.060*"bitch" + 0.059*"friendly" + 0.056*"act" + 0.042*"feed" + 0.039*"dead"

70	0.139*"imagine" + 0.097*"remove" + 0.086*"several" + 0.086*"snack" + 0.069*"difficult" + 0.059*"deep" + 0.055*"energy" + 0.054*"tissue" + 0.048*"guest" + 0.044*"welcome"
71	0.577*"take" + 0.126*"med" + 0.032*"onion" + 0.028*"help" + 0.023*"time" + 0.020*"almond" + 0.017*"need" + 0.017*"prescribe" + 0.016*"therapy" + 0.013*"ridiculous"
72	0.291*"guess" + 0.167*"post" + 0.063*"produce" + 0.057*"version" + 0.055*"substitute" + 0.050*"potentially" + 0.049*"inside" + 0.046*"bean" + 0.034*"fatal" + 0.029*"could"
73	0.312*"eye" + 0.214*"whole" + 0.105*"burn" + 0.068*"whole_life" + 0.036*"forever" + 0.034*"watery_eye" + 0.025*"kitty" + 0.022*"lash" + 0.021*"watery" + 0.020*"hang"
74	0.129*"single" + 0.101*"suddenly" + 0.083*"dose" + 0.048*"nasty" + 0.047*"tooth" + 0.041*"mast_cell" + 0.040*"histamine" + 0.039*"substance" + 0.036*"coat" + 0.033*"immune_response"
75	0.393*"body" + 0.200*"full" + 0.091*"cancer" + 0.089*"fight" + 0.042*"cell" + 0.032*"nature" + 0.026*"boss" + 0.022*"can_breathe" + 0.014*"cause" + 0.012*"son"
76	0.109*"tip" + 0.094*"hell" + 0.084*"stick" + 0.078*"random" + 0.073*"respiratory" + 0.069*"double" + 0.062*"depression" + 0.045*"somewhere" + 0.041*"evening" + 0.028*"pine"
77	0.494*"say" + 0.137*"show" + 0.035*"group" + 0.034*"mental_health" + 0.031*"piece" + 0.028*"mental" + 0.025*"people" + 0.024*"study_show" + 0.017*"shame" + 0.014*"officially"
78	0.107*"open" + 0.078*"apple" + 0.072*"cost" + 0.070*"latex" + 0.064*"brain" + 0.053*"depend" + 0.052*"period" + 0.051*"interesting" + 0.048*"adopt" + 0.044*"receive"
79	0.684*"tell" + 0.091*"tonight" + 0.051*"mild" + 0.038*"go" + 0.028*"people" + 0.022*"give" + 0.019*"time" + 0.015*"could" + 0.013*"saliva" + 0.010*"actually"
80	0.449*"would" + 0.071*"like" + 0.069*"hospital" + 0.061*"name" + 0.050*"would_like" + 0.046*"mind" + 0.032*"sort" + 0.031*"could" + 0.021*"sandwich" + 0.020*"give"
81	0.306*"little" + 0.181*"girl" + 0.180*"pill" + 0.059*"corn" + 0.035*"harm" + 0.028*"fly" + 0.022*"airline" + 0.017*"participate" + 0.016*"give" + 0.016*"help"
82	0.197*"order" + 0.096*"cheese" + 0.073*"matter" + 0.062*"fast" + 0.055*"regular" + 0.050*"major" + 0.050*"clearly" + 0.041*"example" + 0.039*"heart_attack" + 0.031*"ruin"
83	0.475*"life" + 0.195*"life_threaten" + 0.081*"stomach" + 0.077*"threaten" + 0.030*"push" + 0.022*"player" + 0.019*"bar" + 0.019*"variety" + 0.017*"opportunity" + 0.013*"ache"
84	0.212*"hurt" + 0.166*"decide" + 0.137*"normal" + 0.136*"anti" + 0.056*"perhaps" + 0.047*"grateful" + 0.038*"datum" + 0.031*"proof" + 0.029*"planet" + 0.020*"reveal"
85	0.247*"fact" + 0.155*"chocolate" + 0.082*"wheat" + 0.065*"sweet" + 0.055*"illness" + 0.053*"cookie" + 0.028*"grain" + 0.027*"chronic_illness" + 0.027*"delicious" + 0.026*"immunity"

86	0.254*"issue" + 0.127*"high" + 0.125*"meat" + 0.075*"bite" + 0.051*"health_issue" + 0.041*"switch" + 0.032*"benefit" + 0.029*"people" + 0.025*"cause" + 0.023*"mistake"
87	0.217*"seem" + 0.156*"kind" + 0.149*"play" + 0.112*"wish" + 0.064*"best" + 0.057*"wish_could" + 0.045*"could" + 0.041*"phone" + 0.029*"photo" + 0.019*"rich"
88	0.320*"keep" + 0.255*"week" + 0.094*"house" + 0.049*"student" + 0.037*"blow" + 0.024*"past_week" + 0.020*"time" + 0.019*"nail" + 0.018*"bath" + 0.016*"go"
89	0.216*"breathe" + 0.145*"seriously" + 0.115*"recommend" + 0.088*"tweet" + 0.074*"constantly" + 0.056*"miserable" + 0.054*"simply" + 0.032*"highly_recommend" + 0.027*"dislike" + 0.022*"bronchitis"
90	0.336*"call" + 0.139*"truth" + 0.072*"amount" + 0.064*"lie" + 0.053*"cheap" + 0.040*"dream" + 0.032*"fully" + 0.028*"veggie" + 0.025*"replace" + 0.022*"give"
91	0.260*"twitt" + 0.148*"do" + 0.098*"later" + 0.080*"consider" + 0.053*"together" + 0.042*"filter" + 0.029*"involve" + 0.028*"vacuum" + 0.026*"deserve" + 0.025*"test_do"
92	0.467*"face" + 0.126*"daughter" + 0.056*"mask" + 0.054*"face_swell" + 0.051*"face_mask" + 0.045*"impact" + 0.044*"community" + 0.031*"carpet" + 0.025*"cleaning" + 0.024*"original"
93	0.280*"live" + 0.116*"yesterday" + 0.110*"visit" + 0.059*"suggest" + 0.047*"aware" + 0.044*"history" + 0.034*"bullshit" + 0.032*"grocery_store" + 0.026*"gross" + 0.025*"tough"
94	0.268*"fall" + 0.255*"cough" + 0.103*"fall_asleep" + 0.055*"clothe" + 0.052*"gold" + 0.048*"asleep" + 0.041*"go_away" + 0.028*"go" + 0.026*"dance" + 0.025*"fabric"
95	0.284*"long" + 0.147*"list" + 0.085*"nuts" + 0.077*"chemical" + 0.071*"glad" + 0.045*"long_term" + 0.042*"term" + 0.038*"time" + 0.034*"list_ingredient" + 0.020*"role"
96	0.199*"reason" + 0.143*"dust" + 0.142*"clean" + 0.074*"blood" + 0.064*"mold" + 0.050*"expensive" + 0.048*"surgery" + 0.040*"blood_pressure" + 0.040*"pressure" + 0.035*"shower"
97	0.325*"skin" + 0.129*"different" + 0.110*"forget" + 0.076*"sensitive" + 0.064*"sensitive_skin" + 0.057*"coffee" + 0.048*"makeup" + 0.029*"use" + 0.016*"charge" + 0.015*"wear_makeup"
98	0.214*"shit" + 0.161*"point" + 0.155*"sometimes" + 0.104*"allow" + 0.093*"anymore" + 0.037*"people" + 0.022*"suspect" + 0.021*"spicy" + 0.020*"health_condition" + 0.020*"age"
99	0.223*"cold" + 0.201*"hate" + 0.124*"fuck" + 0.084*"fucking" + 0.083*"spend" + 0.032*"party" + 0.028*"garlic" + 0.021*"factor" + 0.018*"climate" + 0.015*"people"
100	0.145*"experience" + 0.145*"certain" + 0.094*"soon" + 0.091*"certain_food" + 0.075*"sensitivity" + 0.057*"food" + 0.048*"occur" + 0.042*"reaction_occur" + 0.037*"soon_eat" + 0.037*"immune_system"
101	0.140*"small" + 0.109*"completely" + 0.085*"line" + 0.078*"alone" + 0.078*"straight" + 0.069*"step" + 0.060*"neck" + 0.043*"trouble" + 0.042*"ear" + 0.039*"plastic"

102	0.309*"first" + 0.187*"first_time" + 0.113*"time" + 0.087*"heart" + 0.069*"likely" + 0.067*"discover" + 0.063*"anxiety" + 0.020*"wind" + 0.015*"knee" + 0.010*"indian"
103	0.260*"water" + 0.082*"severely" + 0.077*"tear" + 0.070*"severely_allergic" + 0.069*"country" + 0.057*"eye_water" + 0.056*"warm" + 0.047*"space" + 0.042*"luckily" + 0.033*"badly"
104	0.258*"away" + 0.204*"stay" + 0.112*"cry" + 0.081*"stay_away" + 0.060*"sell" + 0.054*"blame" + 0.029*"approach" + 0.027*"interested" + 0.026*"cooking" + 0.019*"success"
105	0.260*"else" + 0.103*"totally" + 0.097*"state" + 0.082*"concern" + 0.067*"business" + 0.044*"will" + 0.043*"vaccinate" + 0.040*"people" + 0.030*"advise" + 0.021*"responsible"
106	0.136*"pollen" + 0.106*"throat" + 0.101*"seasonal" + 0.077*"offer" + 0.069*"grass" + 0.055*"sore_throat" + 0.053*"spring" + 0.048*"allergist" + 0.046*"sore" + 0.037*"twitter"
107	0.481*"cat" + 0.068*"touch" + 0.060*"highly" + 0.052*"miss" + 0.046*"highly_allergic" + 0.045*"date" + 0.044*"available" + 0.027*"practice" + 0.020*"stock" + 0.020*"twitter"
108	0.435*"never" + 0.139*"hear" + 0.048*"flight" + 0.046*"multiple" + 0.028*"foodallergy" + 0.028*"info" + 0.028*"complain" + 0.023*"know" + 0.020*"people" + 0.018*"knock"
109	0.300*"home" + 0.111*"instead" + 0.082*"favorite" + 0.062*"flower" + 0.059*"fire" + 0.047*"ready" + 0.047*"fill" + 0.041*"challenge" + 0.036*"twitter" + 0.036*"confirm"
110	0.205*"second" + 0.155*"mother" + 0.148*"bee" + 0.143*"exactly" + 0.049*"okay" + 0.038*"peace" + 0.036*"wise" + 0.031*"curse" + 0.027*"terribly" + 0.025*"time"
111	0.169*"daily" + 0.120*"heat" + 0.096*"expose" + 0.095*"price" + 0.091*"common_sense" + 0.064*"adhesive" + 0.061*"ban" + 0.055*"bump" + 0.041*"common" + 0.031*"sense"
112	0.252*"doctor" + 0.155*"vaccine" + 0.131*"see" + 0.105*"suck" + 0.061*"shoot" + 0.048*"specific" + 0.046*"continue" + 0.035*"lucky" + 0.033*"afraid" + 0.028*"people"
113	0.295*"leave" + 0.103*"terrible" + 0.082*"manage" + 0.076*"perfume" + 0.067*"everywhere" + 0.050*"seed" + 0.032*"roommate" + 0.031*"bill" + 0.022*"employee" + 0.021*"drink_coffee"
114	0.156*"last_week" + 0.107*"game" + 0.091*"speak" + 0.078*"note" + 0.065*"week" + 0.060*"trump" + 0.056*"last" + 0.046*"stream" + 0.045*"join" + 0.040*"fail"
115	0.218*"diet" + 0.113*"bread" + 0.072*"slightly" + 0.072*"vegetarian" + 0.069*"boyfriend" + 0.062*"way" + 0.054*"window" + 0.046*"rice" + 0.036*"yeast" + 0.032*"goal"
116	0.228*"real" + 0.150*"weird" + 0.088*"chronic" + 0.087*"half" + 0.045*"power" + 0.044*"trick" + 0.036*"personally" + 0.033*"thankfully" + 0.028*"gain" + 0.025*"thing"
117	0.151*"other" + 0.092*"candy" + 0.091*"currently" + 0.085*"lunch" + 0.073*"dangerous" + 0.053*"rule" + 0.041*"excited" + 0.037*"ste" + 0.036*"incredibly" + 0.031*"shampoo"

118	0.532*"food" + 0.120*"child" + 0.063*"people" + 0.049*"twitter" + 0.044*"many" + 0.039*"idea" + 0.018*"know" + 0.016*"give" + 0.016*"could" + 0.015*"foodallergie"
119	0.281*"tree" + 0.235*"nut" + 0.094*"tree_nut" + 0.087*"catch" + 0.082*"refuse" + 0.036*"awesome" + 0.030*"catch_cold" + 0.029*"conversation" + 0.027*"walnut" + 0.023*"deliver"
120	0.224*"almost" + 0.197*"die" + 0.064*"almost_die" + 0.049*"event" + 0.047*"lady" + 0.040*"market" + 0.034*"breed" + 0.033*"oral" + 0.032*"burger" + 0.027*"haha"
121	0.331*"much" + 0.145*"sleep" + 0.122*"medicine" + 0.061*"realize" + 0.055*"pretty_much" + 0.032*"protect" + 0.027*"pretty" + 0.025*"scary" + 0.023*"tiny" + 0.021*"migraine"
122	0.202*"remember" + 0.172*"deal" + 0.113*"weather" + 0.109*"fake" + 0.051*"front" + 0.041*"salad" + 0.035*"time" + 0.034*"decade" + 0.032*"waste" + 0.028*"package"
123	0.142*"rash" + 0.111*"poor" + 0.102*"suppose" + 0.071*"create" + 0.067*"light" + 0.065*"look_forward" + 0.053*"indoor" + 0.037*"alive" + 0.036*"forward" + 0.032*"wipe"
124	0.128*"contact" + 0.116*"move" + 0.099*"contact_dermatitis" + 0.092*"support" + 0.067*"number" + 0.059*"extra" + 0.042*"dermatitis" + 0.039*"sadly" + 0.036*"blue" + 0.033*"quick"
125	0.337*"day" + 0.068*"weekend" + 0.051*"couple_day" + 0.049*"barely" + 0.043*"induce" + 0.039*"social" + 0.037*"heal" + 0.037*"medium" + 0.035*"social_medium" + 0.032*"various"
126	0.416*"today" + 0.161*"hour" + 0.105*"wake" + 0.034*"appointment" + 0.034*"twitter" + 0.032*"go" + 0.027*"table" + 0.027*"online" + 0.026*"accommodate" + 0.023*"schedule"
127	0.221*"fish" + 0.182*"story" + 0.158*"throw" + 0.099*"beautiful" + 0.070*"comment" + 0.044*"wild" + 0.033*"joint" + 0.032*"negative" + 0.026*"airborne" + 0.021*"people"
128	0.350*"great" + 0.148*"itch" + 0.093*"team" + 0.070*"moment" + 0.048*"chest" + 0.031*"recover" + 0.022*"prior" + 0.021*"mix" + 0.019*"make" + 0.018*"either"
129	0.151*"base" + 0.115*"tired" + 0.112*"book" + 0.096*"perfect" + 0.092*"plant_base" + 0.064*"middle" + 0.038*"able_breathe" + 0.034*"ambulance" + 0.033*"plant" + 0.032*"straw"
130	0.456*"thank" + 0.113*"share" + 0.095*"follow" + 0.034*"preference" + 0.031*"bleed" + 0.029*"quickly" + 0.026*"apartment" + 0.024*"help" + 0.024*"twitter" + 0.022*"thinking"
131	0.134*"link" + 0.126*"article" + 0.088*"lovely" + 0.066*"focus" + 0.059*"passenger" + 0.050*"physically" + 0.049*"absolute" + 0.043*"twitter" + 0.038*"delay" + 0.036*"click_link"
132	0.325*"dog" + 0.152*"attack" + 0.092*"asthma_attack" + 0.089*"cat_dog" + 0.056*"require" + 0.046*"add" + 0.041*"outside" + 0.029*"individual" + 0.019*"wrap" + 0.018*"leader"
133	0.208*"read" + 0.083*"however" + 0.077*"serve" + 0.066*"customer" + 0.066*"alternative" + 0.050*"train" + 0.049*"prepare" + 0.036*"formula" + 0.033*"introduce" + 0.029*"vaccination"

134	0.175*"talk" + 0.102*"smoke" + 0.099*"possible" + 0.093*"human" + 0.089*"guy" + 0.046*"people" + 0.046*"dumb" + 0.044*"lung" + 0.032*"kinda" + 0.030*"policy"
135	0.286*"hard" + 0.176*"intolerance" + 0.083*"lactose_intolerance" + 0.065*"menu" + 0.035*"pillow" + 0.032*"find" + 0.031*"lactose" + 0.031*"wool" + 0.029*"worker" + 0.027*"unable"
136	0.154*"summer" + 0.137*"research" + 0.126*"penicillin" + 0.121*"winter" + 0.049*"accord" + 0.039*"action" + 0.035*"site" + 0.034*"hungry" + 0.032*"spring_summer" + 0.029*"road"
137	0.147*"worth" + 0.132*"hayfever" + 0.131*"wife" + 0.081*"year_round" + 0.072*"round" + 0.069*"freak" + 0.052*"correct" + 0.045*"thankful" + 0.042*"wheeze" + 0.037*"regardless"
138	0.317*"ingredient" + 0.056*"ingredient_list" + 0.050*"adverse_reaction" + 0.048*"beat" + 0.046*"clinic" + 0.037*"powder" + 0.032*"twitter" + 0.029*"adverse" + 0.027*"technically" + 0.026*"probiotic"
139	0.297*"milk" + 0.218*"dairy" + 0.214*"drink" + 0.043*"exist" + 0.042*"dairy_free" + 0.032*"eczema" + 0.017*"smoking" + 0.014*"make" + 0.013*"give" + 0.012*"find"
140	0.311*"come" + 0.188*"back" + 0.174*"maybe" + 0.081*"come_back" + 0.078*"drug" + 0.045*"thought" + 0.024*"time" + 0.016*"hour_later" + 0.015*"commercial" + 0.011*"give"
141	0.213*"lose" + 0.178*"black" + 0.091*"weight" + 0.064*"patch" + 0.055*"patch_test" + 0.055*"lose_weight" + 0.052*"organic" + 0.044*"remedy" + 0.041*"naturally" + 0.035*"prone"
142	0.652*"allergic_reaction" + 0.272*"reaction" + 0.016*"cause" + 0.009*"give" + 0.008*"twitter" + 0.005*"go" + 0.004*"use" + 0.004*"time" + 0.004*"know" + 0.003*"lavender"
143	0.329*"product" + 0.272*"animal" + 0.104*"choice" + 0.064*"people" + 0.030*"acid" + 0.025*"society" + 0.021*"beauty" + 0.020*"many" + 0.013*"reflux" + 0.013*"animal_product"
	0.239*"mean" + 0.129*"infection" + 0.098*"sinus" + 0.098*"world" + 0.089*"antibiotic" + 0.067*"sinus_infection" + 0.057*"swollen" + 0.041*"breath" + 0.031*"breathing" + 0.026*"sudden"
145	0.189*"tomorrow" + 0.142*"plant" + 0.124*"area" + 0.097*"insurance" + 0.072*"green" + 0.051*"attempt" + 0.046*"food_poisoning" + 0.038*"idiot" + 0.035*"pair" + 0.029*"poisoning"
146	0.253*"free" + 0.206*"gluten" + 0.191*"gluten_free" + 0.132*"pain" + 0.047*"store" + 0.035*"rhinitis" + 0.025*"allergic_rhinitis" + 0.017*"congestion" + 0.015*"twitter" + 0.014*"please"
147	0.132*"condition" + 0.128*"true" + 0.082*"public" + 0.071*"blood_test" + 0.061*"nurse" + 0.059*"attention" + 0.056*"voice" + 0.055*"prove" + 0.034*"supplement" + 0.033*"digestive"
148	0.147*"spray" + 0.112*"nasal_spray" + 0.102*"nasal" + 0.088*"deadly" + 0.080*"otherwise" + 0.073*"rate" + 0.072*"nearly" + 0.052*"inform" + 0.043*"smile" + 0.038*"board"
149	0.193*"watch" + 0.134*"smell" + 0.093*"late" + 0.067*"video" + 0.053*"none" + 0.052*"fresh" + 0.044*"possibly" + 0.040*"meet" + 0.036*"autoimmune_disease" + 0.033*"normally"

150	0.363*"develop" + 0.149*"make_sense" + 0.107*"cool" + 0.085*"sense" + 0.077*"make" + 0.057*"be" + 0.040*"wasp" + 0.024*"cracker" + 0.020*"obesity" + 0.015*"waiter"
151	0.289*"family" + 0.149*"least" + 0.139*"especially" + 0.089*"holiday" + 0.072*"class" + 0.065*"scratch" + 0.020*"left" + 0.018*"kitchen" + 0.017*"deny" + 0.014*"nightmare"
152	0.159*"person" + 0.154*"bring" + 0.125*"place" + 0.111*"wear" + 0.099*"avoid" + 0.083*"buy" + 0.036*"shop" + 0.026*"people" + 0.023*"apply" + 0.018*"arm"

Table 3 Results of LIWC

	Category	Mean	SD
Word Count	**WC**	22.70574	14.36001
Summary Variable	**Analytic**	55.45977	35.33931
	Clout	42.35765	33.31115
	Authentic	45.57744	39.46554
	Tone	34.44495	35.11576
Language Metrics	**WPS**	13.40248	8.668135
	Sixltr	20.89922	13.15111
	Dic	75.76623	15.03119
Function Words	**function**	41.12445	16.12391
Total pronouns	**pronoun**	12.09272	8.813799
Personal pronouns	**ppron**	8.027404	7.11583
	i	5.734562	6.661496
	we	0.321506	1.411346
	you	1.36034	3.402619
	shehe	0.339585	1.771655
	they	0.271803	1.289715
Impersonal pronouns	**ipron**	4.061865	5.285281
Articles	**article**	3.941417	4.696947
Prepositions	**prep**	9.446716	6.939628
Auxiliary verbs	**auxverb**	8.858254	7.395666
Common adverbs	**adverb**	4.682323	5.920813
Conjunctions	**conj**	4.534994	5.085766
Negations	**negate**	1.754165	3.525253
Grammar Other	**verb**	15.40608	9.744432
	adj	3.812438	5.140184
	compare	1.681123	3.341502
	interrog	1.122184	2.77344
	number	1.714235	3.551627
	quant	1.420567	2.847311
Affect Words	**affect**	6.333624	7.117669
	posemo	2.65331	4.633139
	negemo	3.651548	5.806926
	anx	0.318232	1.4137
	anger	1.502514	4.3822
	sad	0.706928	2.360344
Social Words	**social**	5.947932	6.78618

	family	0.228557	1.302946
	friend	0.354516	1.756534
	female	0.279614	1.583418
	male	0.434752	2.071194
Cognitive Processes	**cogproc**	8.985555	8.075405
	insight	1.482487	2.998518
	cause	1.565493	3.113253
	discrep	1.129411	2.802207
	tentat	2.115604	3.936936
	certain	1.15729	2.805092
	differ	2.838586	4.621753
Perpetual Processes	**percept**	2.928496	4.58236
	see	1.206601	2.815288
	hear	0.334947	1.510791
	feel	0.967067	2.691106
Biological Processes	**bio**	8.439591	8.084892
	body	1.629591	3.710609
	health	5.991296	6.615233
	sexual	0.470662	2.546651
	ingest	0.495388	1.900555
Core Drives and Needs	**drives**	6.524635	6.446847
	affiliation	1.940542	3.61071
	achieve	0.768243	2.209988
	power	2.020599	3.635985
	reward	1.467402	3.299017
	risk	0.812463	2.394235
Time Orientation	**focuspast**	2.629677	4.635333
	focuspresent	11.57177	8.673429
	focusfuture	0.993956	2.696392
Relativity	**relativ**	11.42389	9.619232
	motion	1.254231	2.929391
	space	4.507443	5.504732
	time	5.757958	6.913193
Personal Concerns	**work**	1.043832	2.54351
	leisure	1.522867	3.035969
	home	0.223457	1.126338
	money	0.289143	1.37953
	relig	0.200856	1.436572
	death	0.413379	2.006824

Informal Speech	**informal**	4.475552	6.53148
	swear	1.034926	3.887016
	netspeak	2.886086	5.025998
	assent	0.303072	1.715611
	nonflu	0.301083	1.76855
	filler	0.03598	0.501567
Punctuation	**AllPunc**	24.35202	22.92006
	Period	7.846008	12.14719
	Comma	2.052992	4.217197
	Colon	0.392291	2.001801
	SemiC	0.041069	0.725614
	QMark	1.249353	5.586208
	Exclam	1.241322	6.718665
	Dash	1.503117	5.616129
	Quote	0.368132	2.495081
	Apostro	2.712823	4.676937
	Parenth	0.484059	2.734567
	OtherP	6.460584	11.81847